Clustering Analysis of Heterogeneous Data Using Hybrid Clustering Algorithms

Harshi Garg[1], Niranjan Lal[2]
[1, 2] Mody University of Science and Technology,Lakshmangarh,Sikar,Rajasthan, India

Abstract- *In the present digital scenario there is an overwhelming increase in data, with the increase in data it is urgent to develop an effective and efficient approach to handle and cluster these data for further analysis. This paper includes hybridizations of various algorithms like K-Means, Particle Swarm Optimization(PSO), Ant Colony Optimization(ACO) for clustering analysis of Heterogeneous Data. Some of the past and ongoing work based on cluster analysis is also discussed in this paper. This paper also include several algorithms based on hybridization of different clustering algorithms to tackle the heterogeneity of data. Dealing with heterogeneous data is a challenging task. The existing hybrid system is not much efficient to deal with such data in an efficient manner. We will describe the fundamental aspects of clustered data and will analyze each methodology by doing comparative study of all existing algorithms.*

Keywords- Clustering, K-Means, PSO, ACO, Hybrid Clustering, Swarm Intelligence,ANN.

I. INTRODUCTION

Clustering or Cluster Analysis is an emerging field of Data Mining that has always attracted researchers to analyze the methodology behind its working. Clustering is a method of grouping a set of objects into different clusters based upon their properties. The objects having similar properties are in the same cluster than to those objects in other clusters. Clustering is an unsupervised learning as we do not have any prior label in the data and no class values in it. As the volume of data is increasing day by day, handling it is a very challenging task especially when the data is heterogeneous i.e. data of different nature. One of the most popular Clustering Algorithm to handle such data is K-Means algorithm. It's easy implementation and good performance makes it popular among other algorithms. The other clustering algorithms are Particle Swarm Optimization(PSO), Ant Colony Optimization(ACO). To improve the accuracy or we can say quality of algorithm, hybridization of algorithms is done.

K-Means algorithm is an iterative algorithm which comes under vector quantization. It follow two steps: Cluster Assignment and Move Centroid. The performance of this algorithm is exclusively based on initial choice of cluster centroids. It also get influenced by distance calculation factor in which we are finding nearest center for each data object using Euclidean distance.

Swarm Intelligence(SI) as it's name suggest is the field of biologically inspired artificial intelligence in which we study the collective behavior of ants, wasps, termites, bees. It raises group intelligence by forming flocks, shoals, schools, swarms and colonies. PSO comes under one of the Swarm Intelligence algorithm which is inspired by the behavior of the fish schooling and bird flocking. It is very easy to implement as there are very less parameters to adjust. It's application areas includes: Fuzzy Control System, Training of Artificial Neural Network(ANN) and other where Genetic Algorithms can be applied.

ACO algorithm is inspired by the ant's behavior, searching for food. This algorithm provides a better solution for the optimization problem. Ants drop pheromones everytime they bring food, and because of that shorter paths are become more stronger, hence optimizing the problem solution.

We have organized this paper as follows. Section 2 explains the previous work donein searching and ranking, Section 3 explains the K-Means Algorithm, Section 4 explains the PSO Algorithm, Section 5 explains the ACO, Section 6 explains the Hybridization of K-Means and PSO, Section 7 explains the Hybridization of PSO and ACO, Section 8 explains the Hybridization of Fuzzy Adaptive Particle Swarm Optimization(FAPSO), ACO and K-Means(FAPSO-ACO-K), Section 9 conclude the paper with future directions.

II. RELATED WORK

Many researchers have done hybridization so that drawbacks of one clustering algorithm can be tackled by the strengths of another algorithm.

Koheriet.al. [1], M. R. Khammar & M. H. Marhaban[2] proposed and improved the K-means algorithm using multiple methods so that loophole of random selection of initial centroids can be easily removed.

Van Der Merwe DW et.al.[3] proposed a very first hybridization of K-Means andPSO where every cluster center of K-Means is considered as a particle of the PSO. Then these cluster centers using PSO and other randomly generated initial particles are optimized to give the resulting cluster centers. The strengths of both the clustering algorithm helps in improving the final result of clustering.

Fun Y et.al. [4] proposed an another hybridization of K-Means and PSO called

Alternative K-Means and PSO (AKPSO). Here instead of Euclidean distance they used an alternative metric. PSO generates the sub-optimal solution and K-Means is used for improving the desired result.

Niknam T et.al. [5] proposed an another hybridization of ACO along with K-Means and Fuzzy Adaptive PSO(FAPSO). Here K-Means fructify the hybridization result of ACO and FAPSO.

Dheebet. al. [6] designed, implemented and evaluated a software using K-Means Clustering Algorithm. This is a image-processing based software used for classification and will automatically detect the plant leaf disease.

Nayak J, Kanungo DP et.al. [7] illustrated recently in his paper about the combination of K-Means, Genetic Algorithm and Improved PSO.

P.S Bradley et.al. [8] proposed an efficient algorithm for refining an initial starting point for a general class of clustering algorithms and this approach is fast as compare to other.

III. K-MEANS ALGORITHM

The well known unsupervised learning algorithm used for solving clustering problems. Here the main idea is to define K clusters, one for each cluster. It is a fast and robust algorithm. This algorithm gives better result when data sets are distinct and are well separated from each other. But this algorithm fails if the data set is non linear or we can say for categorical data, this algorithm fails. It can't handle noisy data. The algorithm for K-Means is:

- **Input**

1. K: number of total clusters
2. S: Data set having n objects

- **Output**

1. Group of K(cluster).

- **Algorithm Flow Chart**

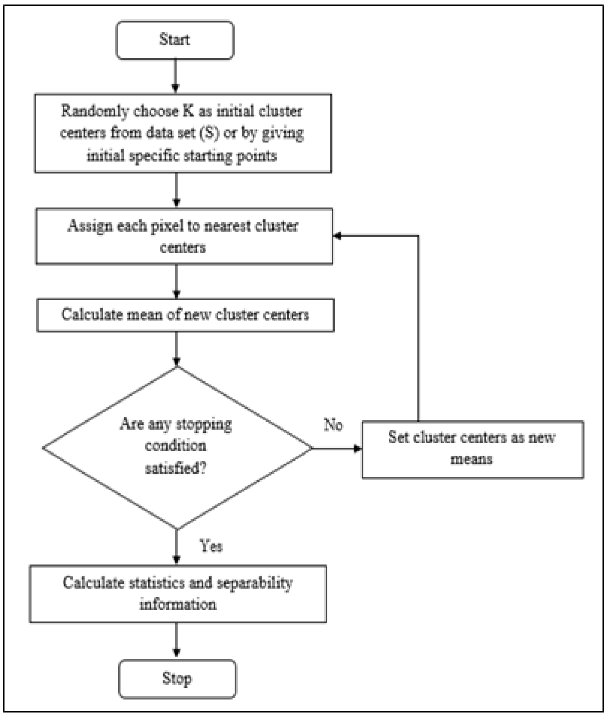
Start
Randomly choose K as initial cluster centers from data set (S) or by giving initial specific starting points
Assign each pixel to nearest cluster centers
Calculate mean of new cluster centers
Are any stopping condition satisfied?
No
Set cluster centers as new means
Yes
Calculate statistics and separability information
Stop

Fig.1 Flow chart of K-Means Algorithm

For choosing an appropriate value of K we can run algorithm with different values of K to obtain better results. Results of K-Mean algorithm are shown below:

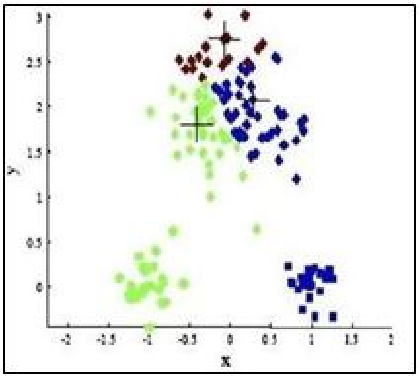

Fig.2 Iteration 0

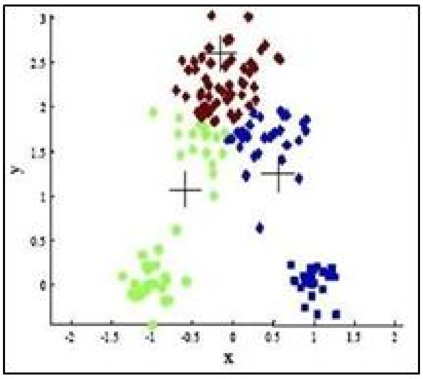

Fig.3 Iteration 1

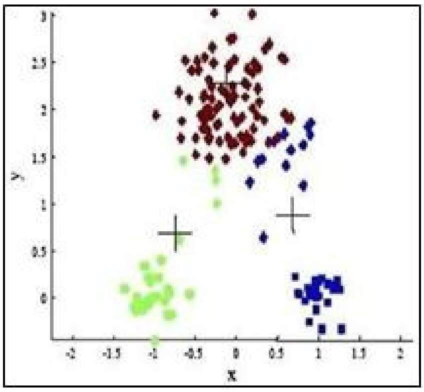

Fig.4 Iteration 2

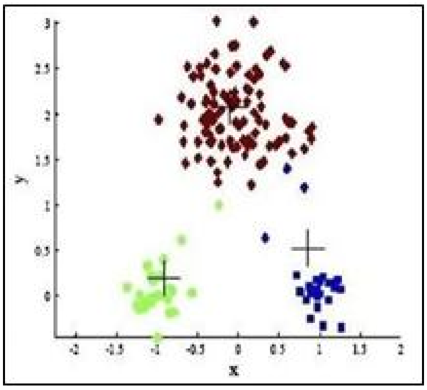

Fig.5 Iteration 3

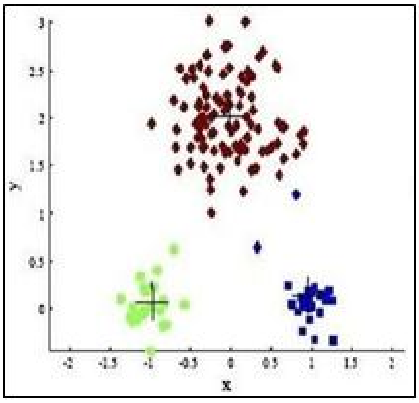

Fig.6 Iteration 4

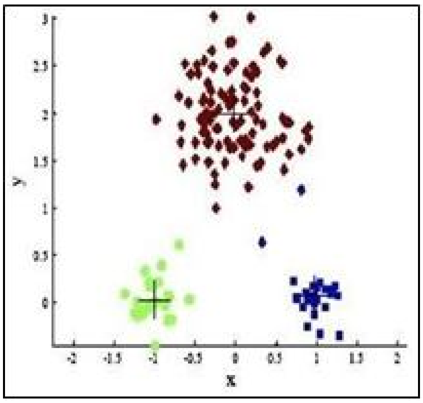

Fig.7 Iteration 5(Convergence)

IV. PARTICLE SWARM OPTIMIZATION

This algorithm is inspired by the behavior of the bird flocking and fish schooling. It is a very simple algorithm which is easy to implement.This algorithm is used for continuous and discrete optimization problems. Suppose group of birds is circling over a certain area where they can smell veiled source of some food. The bird who is closest to the food, tweets the loudest and the other birds whirl will swing around his direction. If any other bird comes closer to the target food than the first bird,it tweets louder and others will swing over towards him.

- This algorithm will keep track of three global variables :

  1. Global Best(gBest) value: It shows which particle's data is now near to the target.
  2. Condition or target value.
  3. Stopping condition to stop the algorithm if the target is not found.

- The particle consist of following:

  1. Personal Best(pBest) value: It shows that the nearest the particle's data hasever come to target.
  2. Velocity value indicating how much data can be changed.
  3. Data representing a possible solution.

- **Algorithm Pseudo code**

```
For each particle
{
Initialization of particle
}
Do Until maximum number of iterations or minimum error criteria
{
```

```
For each particle
{
Calculate it's data Fitness value (Fvalue) If Fvalue is better than pBest
{
Set pBest = current Fvalue
}
If pBest is greater than(>) gBest
{
Set gBest = pBest
}
}
For each particle
{
Calculate particle velocity
gBest and velocity is used to update particle data
}
}
```

Particle swarm Optimization algorithm results are shown in Fig.7:

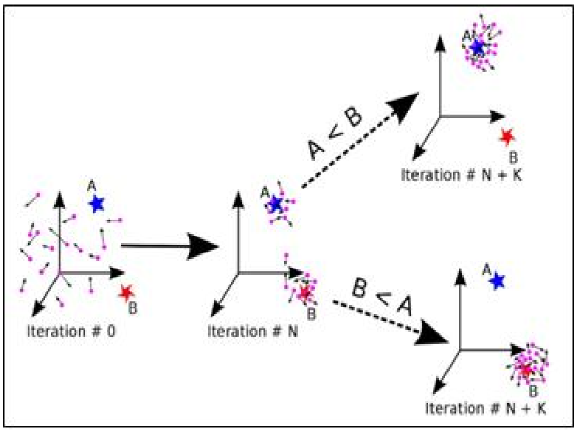

Fig.8 Iterations in PSO

V. ANT COLONY OPTIMIZATION

This algorithm is inspired by the behavior of some ant species. It is population based meta-heuristic optimization. It is efficient for Travelling Salesman problem as shown in Fig. 8 and also used for scheduling purpose. It uses pseudorandom proportional rule. Ants drop the pheromone trail everytime when they bring food and with the help of this, shorter path gets stronger and thus we get the optimal solution. The algorithm pseudo code for the ACO system is:

```
1. Initialize pheromone trail
2. Do While(Stopping Criteria is not satisfied)

   -Do Until (each and every ant completes a tour)
   -Local Trail update
   -End Do
   -Analyze Tours
   -Global Trail update

3. End Do
```

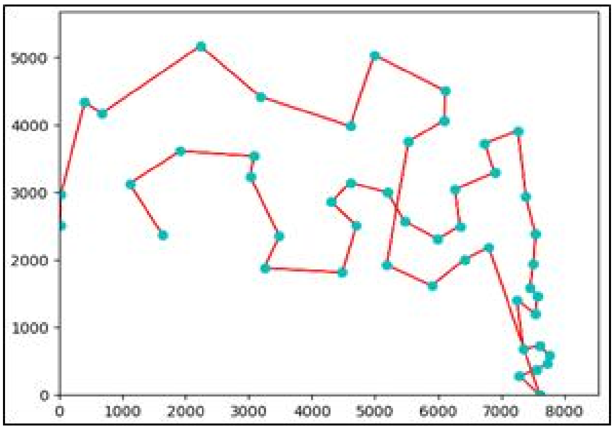

Fig.9 Travelling Salesman problem using ACO

VI. HYBRIDIZATION OF K-MEANS AND PSO

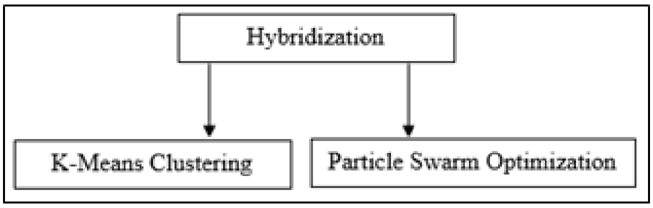

Fig.10 K-Means and PSO Hybridization

K-Means algorithm converges faster than PSO because of fewer function evaluations. But when initial cluster centers choice has a huge effect on result, the accuracy of K-Means algorithm decreases[10]. In PSO based clustering algorithm, particle encoding uses the method of cluster center.

Here each particle Cij= (Ci1, Ci2,…….CiK) indicates the K cluster centered vectors, Cij refers here to the jth cluster centered vector of the ith particle. The vector Zij= (Zi1, Zi2,……, ZiS) indicates the position of ith cluster of each particle in cluster, S is the number of attributes of sample.

The K-Means + PSO algorithm procedure is defined as follows:

1. Initialization: Set the parameters. Initialize or assign the velocities, positions of a group and computing fitness;
2. Update the velocities and positions of a group;
3. For all particle of new generation, use the k-means clustering algorithm to finish the clustering task;

4. Compute each particle's fitness value and update the gbest and the pbest.
5. If the stopping condition is reached, output the classification outcome, else, go to 2.

There are various variant algorithm present of K-Means and PSO.

- Variants of K-Means + PSO

1. Rough K-Means + PSO
2. K-Harmonic Means + PSO

- K-Means + Variants of PSO

1. K-Means + QPSO(Quantum-Behaved Particle Swarm Optimization)
2. K-Means + PPO(Particle-Pair Optimizer)

VII. HYBRIDIZATION OF PSO AND ACO

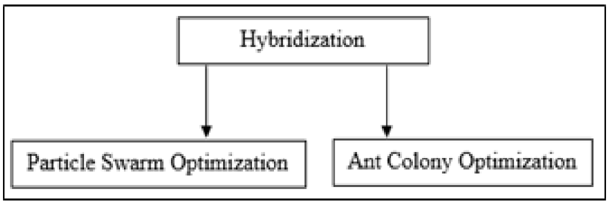

Fig.11 PSO and ACO Hybridization

In PSO, based on the current position, new position of particle is generated. In ACO, based on the pheromone table new ants are generated[9]. The two clustering algorithms exchanges the best solution. This hybridization approach maintains original characteristics of both the clustering algorithm. The algorithm can be summarized as:

1. Randomly initialze the particles;
2. For each particle fitness value is calculated and Best is updated. Best is updated with the best fitness value;
3. For all ants pheromone is calculated;
4. For each ant fitness is calculated;
5. Best-Ant with Best fitness value is updated;
6. Global-opt with the best value of gBest and Best-Ant is updated;
7. Until stopping criteria is true, Do;
8. Generate the new positions of particles;
9. For each particle i=1,2,……N , Do;
10. Calculate the particle fitness and update it's pBest (if fitness of that particle is better than pBest in previous iteration);
11. gBest updation is done (if fitness value of particle is better than gBest) END Do;
12. Generate the new ants with N rows;
13. For each ant i=1,2,……N , Do;
14. Ant fitness is calculated;
15. If fitness value of ant is better than the previous solution, then worst solution is replaced;
16. If fitness value of ant is better than Best-Ant, then Best –Ant is updated; END Do;
17. gBest is exchanged as: worst ant is replaced with the gBest and worst particle is replaced by Best-Ant;
18. Update the Global-opt with the best value of Best-Ant and gBest; END Do;
19. Return Global-opt, which is the final optimal solution.

VIII. HYBRIDIZATION OF FUZZY ADAPTIVE PARTICLE SWARM OPTIMIZATION(FAPSO), ACO AND K-MEANS(FAPSO-ACO-K)

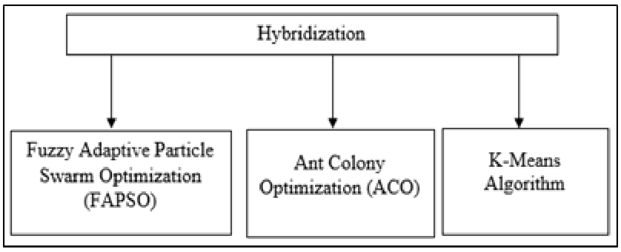
Hybridization
Fuzzy Adaptive Particle Swarm Optimization (FAPSO)
Ant Colony Optimization (ACO)
K-Means Algorithm

Fig.12 FAPSO-ACO-K Hybridization

K-Means Algorithm is not able to handle non-linear data. But PSO is efficient enough to handle non-linear optimization problems. To adapt intertia weight of PSO, fuzzy system is implemented. This algorithm uses the result of PSO-ACO as an initial condition of K-Means. The working method behind this algorithm is: selsction of gBest depends on the ACO best path selection. The algorithm pseudo code can be summarized as:

1. Generate initial population and initial velocity for each and every particle;
2. Generate trail intensity(initial);
3. Calculate objective function value for each individual;
4. Based on objective function values, sort the initial population;
5. Select gBest (depends on individual having minimum objective function);
6. Select pBest (best local position for each individual);
7. Select ith individual;
8. Calculate next position for ith individual;

9. If all individuals are selected then goto next step else i=i+1and goback to step7;
10. Check the stopping criteria(If current iteration number reaches the maximum iteration number, then goto next step, else replace initial population with the new population of swarms and then go back to step 3);
11. Consider gBest value(last) as initial solution for k-means algorithm.

IX. CONCLUSION

Cluster analysis is a vast research area of data mining having number of clustering algorithms used for optimization problem. K-Means is an effective algorithm for clustering large datasets and used in commercial applications. Improvement in PSO algorithm is still ongoing and still there are many application areas in PSO which are unknown like mathematical validation. PSO can also be applied with Genetic Algorithm and Artificial Neural Network. ACO works on a very dynamic system so it is better to use it in graphs with changing topologies like we can use it in computer networks. Hybridization of various clustering algorithms is shown in this paper so that one algorithm can overcome the drawbacks of the another algorithm. In my future work

I will apply hybrid clustering on heterogeneous data using variants of K-Means along with both PSO and ACO and then will find the difference in the results.

REFERENCES

[1] Koheiet. al, "Hierarchical K-means: an algorithm for centroids initialization for K-means", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[2] Khammar, Marhaban, "Obtaining the initial centroids based on the most dense colonies in the k-means Algorithm", Reasearch Journal of Computer Systems & Engineering, ISSN: 2230-8563, Vol. 03, Issue. 01, July 2012.

[3] Van Der Merwe DW, Engelbrecht AP, "Data clustering using particle swarm optimization", In: IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia, pp. 215-220.

[4] Fun Y, Ching-Yi C, " Alternative KPSO-Clustering Algorithm", Tamkang Journal of Science and Engineering, 2005.

[5] Niknam T, Amiri B, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis", Appl Soft Comput 2010;10:183-97.

[6] Dheeb, Malik & Sulieman, "Detection and Classification of Leaf Diseases using K- means-based Segmentation and Neuralnetworks-based Classification", Information Technology Journal, Vol. 10, Issue. 2, pp. 267-275, 2011.

[7] Nayak J, Kanungo DP, Naik B, Behera HS, "Evolutionary Improved Swarm-Based Hybrid K-Means Algorithm for Cluster Analysis", In: Second International Conference on Computer and Communication Technologies: 2016; Springer India. p. 343-52.

[8] Bradley & Fayyad, "Refining Initial Points for K-means Clustering", International Conference of Machine Learning", pp. 91-99, May 1998.

[9] Cheng-LungHuang, Wen-ChenHuang, HungYiChang, Yi-ChunYeh, and Cheng-Yi Tsai, "Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering", Applied Soft Computing, Vol. 13, pp-3864– 3872,2013.

[10] Hai Shen, Li Jin, Yunlong Zhu, Zhu Zhu, "Hybridization of particle swarm optimization with the K-Means algorithm for clustering analysis", In: IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC- TA), Nov. 2010.