IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

A Comparative Study on K-Means and Genetic Algorithm

P. Dheivanai[1*], P. Sundari[2]

[1,2]Department of Computer Science, National College, Trichy, India

[*]*Corresponding Author: deivaddd@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** Data mining is the process of analyze data from different perspectives and summarizing it into useful information. Clustering is a practical unsupervised data mining task that subdivides an input data set into a desired number of subgroups so that members will have high similarity and the member of different groups have large differences. K-means is a usually used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represent by their centroids, by minimizing the square error function. Although K-means is easy and can be used for a wide variety of data types, it is rather sensitive to initial positions of cluster centers. There are 2 simple approach to cluster center initialization i.e. either to select the initial values at random, or to select the first k samples of the data points. Both approach cause the algorithm to converge to sub optimal solutions. Genetic algorithm one of the usually used evolutionary algorithms performs global search to find the solution to a clustering problem. The techniques typically starts with a set of randomly generated individuals called the population and creates successive, latest generations of the population by genetic operations such as natural selection, crossover, and mutation. Each one chromosome of the population represent K no. of centroids. Steps of genetic algorithm are repeatedly applied for a no. of generations to search for suitable cluster centers in the feature space such that a similarity metric of the resultant clusters is optimized. K-means and genetic algorithm based data clustering have been compared in this paper on the basis of their functioning principle, advantage and disadvantage with proper example.

***Keywords*—** Data mining, K-means, Genetic algorithm

I. INTRODUCTION

Data Mining is a knowledge mining process. It is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in bulky data sets involving method at the intersection of artificial intelligence, machine learning, statistics and database systems. The tremendous growth of scientific databases put a lot of challenges before the research to extract useful information from them using traditional data base techniques. Hence effective mining methods are important to discover the implicit information from huge databases. Cluster analysis is one of the major data mining algorithms, extensively used for a lot of practical application in various emerging areas like Bioinformatics. Clustering is an unverified method that subdivides an input data set into a desired number of subgroups so that the objects of the same subgroup will be similar or associated to one another and different from or unrelated to the objects in other groups. A high-quality clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The excellence of a clustering result depends on both the similarity measure used by the method and its execution and also by its ability to discover Some or all of the hidden patterns K-means is a usually used partitioning based clustering method that tries to find a user specified number of clusters (k), which are represent by their centroids, by minimizing the square error function. The clustering method aims at optimizing the cost purpose to minimize the difference of the objects within each cluster, while maximizing the dissimilarity of different clusters.

Genetic Algorithm (GA) parallel search method, that searches for a global approximate solution to the clustering troubles through application of the principles of evolutionary biology. The algorithm typically starts with a set of at random chosen solution called the population and creates successive, new generations of the population by genetic operations such as natural selection, crossover, and mutation. Natural selection is performed based on the fitness of an individual. An individual, the better its fitness, and the more chances it has to survive in the next generation. Crossover is performing by certain crossover rule and mutation aims at changing an individual by a user-specified mutation prospect. The perception fundamental the approach is that each new population will be better than the previous one. Usually, solutions are represented using fixed length strings, particularly binary strings, but alternative encodings have been developed. The greatest advantage of genetic algorithms is that the fitness function can be altered to change the behavior of the algorithm. The next section provides a detailed over view of K-Means based data

clustering with proper example. The section 3 contains the details of K-Means and GA based data clustering with proper example. A comparative analysis has specified in section 4 followed by conclusion.

II. LITERATURE SURVEY

Agustin Blas et al.[8] explains the grouping Genetic algorithm in clustering performance, started with proposed encoding, and different modification of crossover and mutation operation and also initiated the local search include with the island model for improve the performance of the problem. The real data sets like iris and wine were used and compared the results with the classical approaches such as DBSCAN and K-means, and obtaining the excellent results in proposed grouping based methodology the evolutionary approach such as Genetic algorithm. The performance of the algorithm was measured by using the different fitness function.

Tzung-Pei-Hong et al.[9] discussed the performance of the Genetic algorithm based attribute clustering process were improved based on the grouping Genetic algorithm. The Genetic operations, chromosome representation and fitness function defined in grouping Genetic algorithm to solve the clustering problem. As the result of grouping Genetic algorithm based clustering algorithm improved the convergence speed and fitness value of the clustering problem. In addition the algorithm can also deal with the problem of missing values. The other optimization algorithms are used to solve the problem in attribute grouping.

Daniel Gomes Ferrari et al. [10] proposed a new approach to characterize the clustering problems based on the similarity among objects and the method for combine internal indices for ranking algorithms based on the performance of the problem. The experimental results indicated the viability of meta learning systems for an unlabeled approach to the clustering algorithm selection problem. This technique presents the better result from the distance based set over the attribute based approach.

Kunnuri Lahari et al. [13] enhanced reduce the local minima using evolutionary and population based methods like Genetic algorithm and teaching learning based optimization. The data sets wine and iris are used, and the experimental results are compared with the Genetic algorithm and teaching learning based optimization based clustering with k-means algorithm. The performance of the evolutionary based clustering method compared with some existing clustering method.

Rahila H.Sheikh et al. [14] proclaimed a brief study of Genetic algorithm based clustering. Rajashree Dash et al.[11] discussed on comparative analysis of K-means and Genetic algorithm based on clustering. Arun Prabha et al. [15] with respect to the idea was improved the cluster quality from K-means clustering using a Genetic algorithm. Large scale clustering problems in data mining also address by this method. The best results are achieved by using this method.

Anusha et al.[16] for optimal clustering depicted an enhanced K-means Genetic algorithm . The author overcomes the Anomalies of local dataset the algorithm fails in computational time and also optima with suitable. It is inferred that the technique produced more than the 90% accuracy for real life dataset. The author also tried a neighborhood knowledge strategy for optimizing multi objective troubles. This algorithm utilize the k means Genetic algorithm for finding the smallness of the clusters. By this results that the algorithm could produce lowest index value for the maximum datasets.

III. METHODOLOGY

A. Overview Of Datamining

Data Mining is the innovation of hidden information found in large quantities of data and can be viewed as a step in the knowledge discovery process (Fayyad 1996).Data mining defined as a set of computer-assisted techniques designed to automatically mine big volumes of integrated data for new, hidden or unexpected information, or interesting patterns. With small set of data, traditional statistical analysis can be efficiently used. The first and simplest analytical step in data mining is to explain the data summarize its statistical attributes such as means and standard deviations, visually evaluation it using chart and graphs, and look for potentially meaningful links among variables such as values that often occur together (Edelstein 1998).

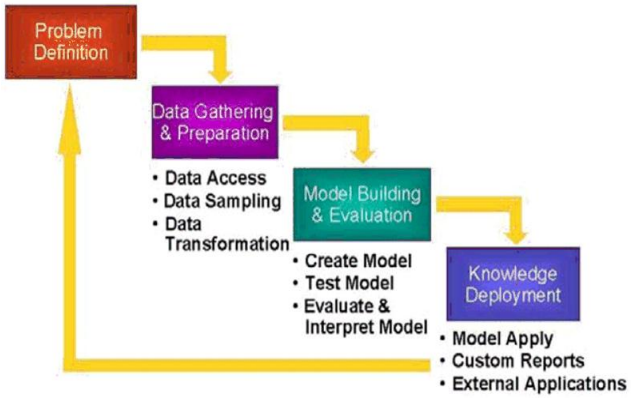

Figure 3.1 An overview of steps that compose KDD process

B. K-Means Algorithm

K-means clustering is a partitioning based clustering method of classifying/grouping items into k groups (where k is user

specified number of clusters). The group is done by minimizing the sum of squared distances (Euclidean distances) between items and the matching centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of standardized density". Although K-means is simple and can be used for a wide variety of data types, it is moderately sensitive to initial position of cluster centers. There are two simple approaches to cluster center initialization i.e. either to choose the initial values randomly, or to choose the first k samples of the data points. As an option, different sets of original values are selected out of the data points and the set, which is closest to optimal, is chosen.

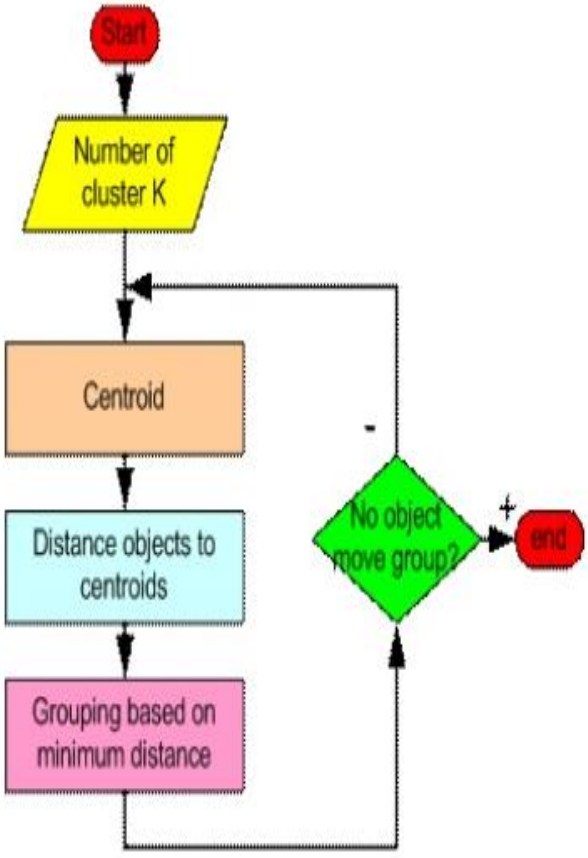

Figure : 3.2 K-Means Clustering Works

C. Genetic Algorithm

Genetic algorithms introduced by John Holland at the University of Michigan in the early 1970's. Genetic algorithms are theoretically and empirically established to provide robust search in complex spaces (Goldberg, 1989). Genetic algorithms are stochastic search method that mimic natural genetic evolution. Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of ordinary selection and genetics.

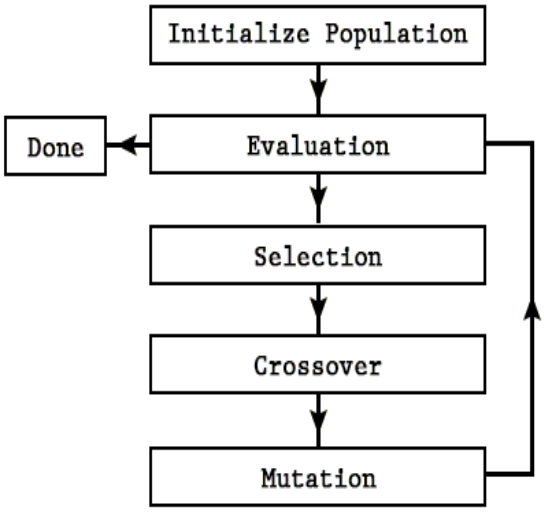

Figure: 3.3 Genetic Algorithms Overview

D. Proposed Algorithms

K-MEANS Algorithm:

K-means cluster is a partition based clustering technique of classifying/grouping items into k groups (where k- is user specified number of clusters).

Algorithm: Original K-means(S, k), S= $\{x_1, x_2 \ldots x_n\}$.

Input: The no of clusters k and a dataset contain n objects x_i.

Output: A set of k clusters C_j that minimize the squared-error criterion.

Begin

1. m=1;
2. initialize k prototypes; //arbitrarily chooses k objects as the initial centers.
3. Repeat

for i=1 to n do

begin

for j=1 to k do

Compute $D(X_i, Z_j) = |X_i, Z_j|$;//$Z_j$is the center of cluster j.

if $D(X_i, Z_j) = \min\{D(X_i, Z_j)\}$ then

$X_i \in C_j$;

end;// (re)assign each object to the cluster based on the mean

if m=1 then

$$J_c(m) = \sum_{j=1}^{k} \sum_{X_i \in C_j} |X_i - Z_j|^2$$

m=m+1;

for j=1 to k do

$$Z_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)}$$; //(re)calculate the mean value of the objectsfor each cluster

$$J_c(m) = \sum_{j=1}^{k} \sum_{X_i \in C_j} |X_i - Z_j|^2$$; //compute the error function

4. Until $J_c(m) - J_c(m-1) < \zeta$

End

Genetic Algorithms

Genetic algorithm facts are as follows:

- Heuristic Search Algorithms Method based on evolutionary ideas of natural selection and genetics Provide efficient, effective algorithms for optimization
- Useful when search space very large or too complex for analytic treatment

Algorithm Key concept are as following:

1. Individual - Any potential solution
2. Genes-Attributes of an entity
3. Population - collection of all *individuals*
4. Search Space - All potential solutions to the trouble
5. Chromosome – (set of genes) plan for an *individual*
6. Fitness function- A function that assign a fitness value to an individual
7. Genetic operator:
   - Reproduction [Selection]
   - Crossover[or Recombination]
   - Mutation-(Altering or Modifying)

The Algorithm GAs

1. randomly initialize population(t)
2. determine fitness of population(t)
3. repeat
   1. select parents from population(t)
   2. perform crossover on parents creating population(t+1)
   3. perform mutation of population(t+1)
   4. determine fitness of population(t+1)
4. until best individual is good enough

IV. EXPERIMENT AND RESULTS

The proposed system has been implemented using .NET environment. It can be executed on windows. The results are obtained as follows after execution. Sample data set is presented.The Main theme of this literature survey is comparative study in k-means & GA. Data a mushroom, Irish, Soybeans dataset has been used with 119 items each for analysis. A set of association rules are obtained by applying K-Means and Genetic Algorithm. By analyzing the data, and giving different support and confidence values, execution time, can obtain different number of rules. During analysis it found that Genetic is much faster for huge number of transactions as compare to K-means. It takes less time to generate frequent item sets. We work on mushroom data which contains 8124 transactions.

Table 4.1 showing comparison of various dataset in the proposed algorithm

Dataset	No of Record	Number of Items	Number of Items Per Record
Mushroom	8124	119	30
Soybean	683	36	36
Fishers Iris	150	10	12

4.2 MEMORY SPACE & EXECUTION TIME OUTPUT

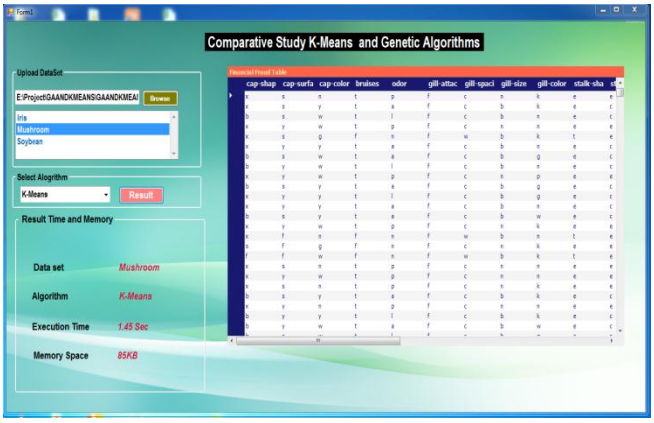

Figure 4.2 (a) .Mushroom Dataset using K-Means Implementation in Time & Memory space

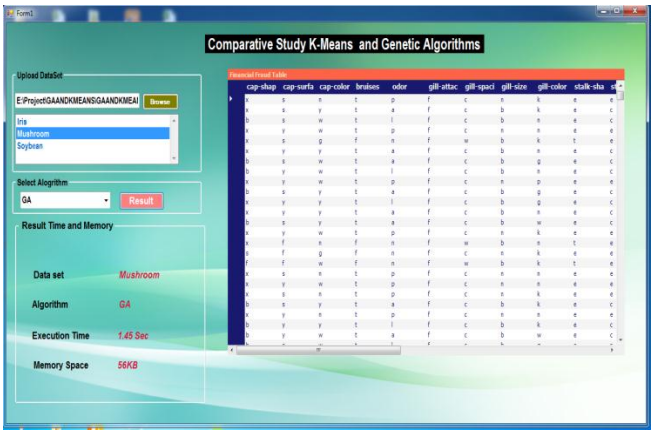

Figure 4.2 (b) .Mushroom Dataset using Genetic Implementation in Time & Memory space

In this study, types of techniques are used are two to find out the memory space, support, confidence, execution time accuracy of mushroom, Irish, soybeans data set. When compared to GA technique, K-Means achieved High accuracy.

Comparison of Various Dataset using Algorithms.

Indicating the support, confidence, memory space, time taken of the 2 methods used in this study is given in Table 2.

Table 4.2(c) showing Minimum Support for all Dataset

Dataset	K-Means	Genetic
	Minimum Support	
Mushroom	2.45	0.45
Soybean	0.17	0.5
Iris	0.2	0.1

Table 4.2(d) showing Confidence for all Dataset

Dataset	K-Means	Genetic
	Confidence	
Mushroom	0.5	0.3
Soybean	0.4	0.2
Iris	0.8	0.5

Table 4.2(e) Showing memory Space for all Dataset

Dataset	K-Means	Genetic
	Memory Space	
Mushroom	85KB	56KB
Soybean	5KB	3KB
Iris	75KB	45KB

Table 4.2(f) showing comparison of Various Dataset using GA & K-Means Algorithms

Dataset	K-Means Time Taken (in mil.secs.)	Genetic Time Taken (in mil.secs.)
Mushroom	2.60	1.45
Soybean	0.25	0.17
Fishers Iris	0.8	0.2

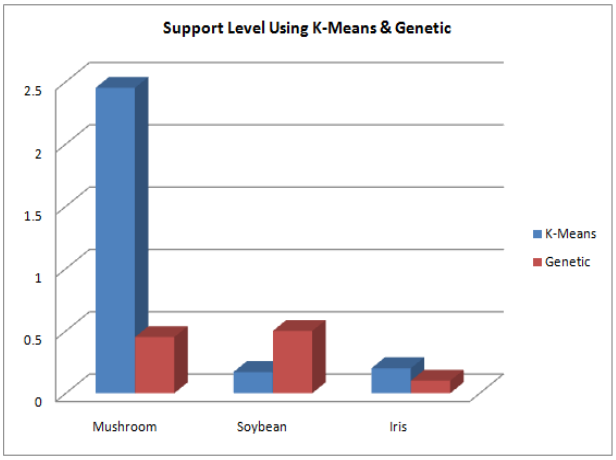

Chart 4.2(c) Graph representing Minimum Support

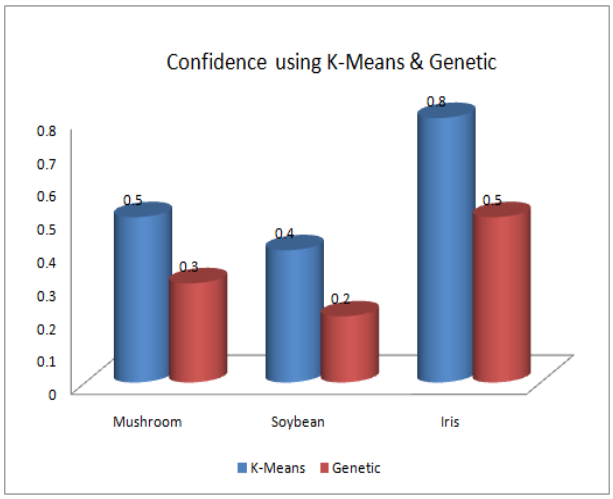

Figure 4.2(d) Graph representing Confidence

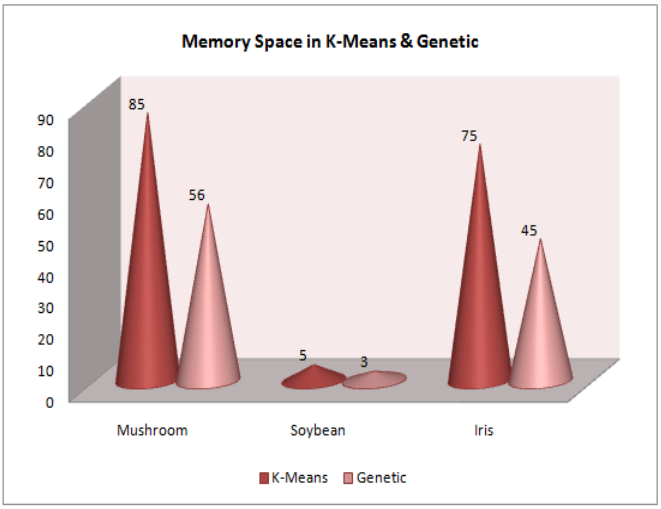

Chart 4.2(e) Graph representing memory Space

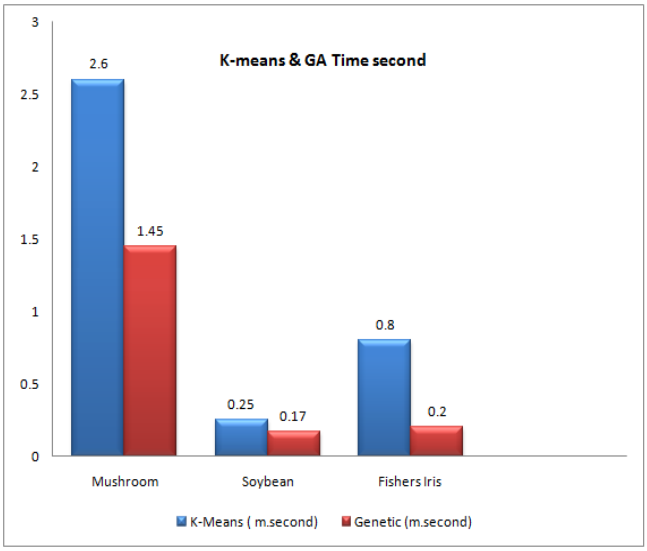

Chart 4.2(f) Graph representing Various Dataset measured in Seconds

Table 4.3 showing Final Comparison of Genetic Algorithm & K-Means

K-MEANS BASED DATA CLUSTERING	GA BASED DATA CLUSTERING
Partitioning Based Method	Evolutionary Based Method
Input: k, dataset, randomly chosen k centroids	Input: k, dataset, P, randomly chosen P chromosomes, tmax,.
Objective: Minimizing sum of squared distance	Objective: Minimizing the sum of distances from each data point to its cluster centroid
Termination condition: No changes in new cluster centroids.	Termination condition: Maximum no. of iterations reached.
Final clustering may converge to local optima.	GA is based on global search approaches with implicit parallelism.
Time complexity: O(n*k*d*i) Where n= no. of data points k= no. of clusters d= dimension of data i= no. of iterations	Time complexity: O(tmax*p*n*k*d) Where n=no. of data points k= no. of clusters d= dimension of data tmax= maximum no. of iterations p= population size

V. CONCLUSION AND FUTURE SCOPE

In this comparative study K-means & GA techniques are used to find out the support, confidence, memory space, time second of Mushroom,Soybean,Fishers Iris data. High accuracy achieved through GA technique compare than K-Means achieved compare than algorithms. Clustering is an important unsupervised classification technique where a set of data objects taken in a multi-dimensional space, are group into clusters in such a way that data objects in the same cluster are parallel in some sense and substance in different clusters are dissimilar in the same sense. K-Means is an intuitively simple and effective clustering technique, but it may get stuck at suboptimal solutions, depending on the choice of the initial cluster centers. Under certain conditions, to provide an optimal clustering is expected in a GA-based clustering technique, more superior to that of K-Means algorithm, but with little more time complexity.

REFERENCES

[1] Nikita Jain,Vishal Srivastava, "Data Mining techniques : A survey paper" , International Journal of Research in Engineerning and Technology, pp. 116-119, 2013.

[2] M.S.B PhridviRaj, C.V. GuruRao, "Data Mining – Past present and future data streams," Elsevier, pp. 256-264, 2013.

[3] K.Kameshwaran, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining," International Journal of Computer Science and Information Technologies, pp.2272-2276, 2014.

[4] Gunjan Verma, Vineeta Verma, "Role and Application of Genetic Algorithm in Data Mining," International Journal of Computer Application, pp. 5-8, 2012.

[5] Sharaf Ansari,Sailendra Chetlur, Srikanth Prabhu, N. Gopalakrishna Kini, Govardhan Hegde, Yusuf Hyder, "An Overview of Clustering Analysis Techniques used in Data Mining ," International Journal of Emerging Technology

[6] Aastha Joshi, Rajneet Kaur, " A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, pp.55-57,2013.

[7] Manoj Kumar, Mohammad Husian, Naveen Upreti, Deepti Gupta, Genetic Algorithm ": Review and Application," International Journal of Information Technology and Knowledge Management, pp.451-454, 2010.

[8] L.E. Agustın-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro- Calvo, J. Del Ser, J.A. Portilla-Figueras K. Elissa, "A new grouping genetic algorithm for clustering problems," Elsevier, pp.9695-9703, 2012.

[9] Honga Tzung-Pei, Chun-Hao Chenc, Feng-Shih Lin, "Using group genetic algorithm to improve performance of attribute clustering," Elsevier, pp.1-8, 2015.

[10] Danial Gomes Ferrari, Leandro Numes de Castro, " Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking combination methods," Elsevier, pp.181-194, 2015.

[11] Rajashree Dash and Rasmita Dash, "Comparative analysis of K-means and Genetic algorithm based data clustering," International Journal of Advanced Computer and Mathematical Sciences, pp.257-265, 2012.

[12] Edvin Aldana-Bobadhilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.

[13] Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Prediction based clustering using genetic algorithm and Learning Based Optimization Performance Analysis," Advances in Intelligent Systems and Computing," pp. 338, 2015.

[14] Rahila H. Sheikh, M. M.Raghuwanshi, Anil N. Jaiswal, "Genetic algorithm based clustering: A Survey," IEEE, pp.314-319, 2008.

[15] K.Arun Prabha, R.Saranya, "Refinement of K-means clustering using Genetic algorithm," Journal of Computer Application, pp. 256-261, 2011.

[16] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE, pp.580-584, 2014.

[17] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.

