International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 12, Issue 11 (November 2016), PP.01-09

A Comparative Study on K-Means And Genetic Algorithm For Data Clustering

[1]Ashok Kumar D, [2]Usha T. A, [3.]Sivaranjani C.
[1,2,3]*Government Arts College, Trichy, Tamilnadu, India*

Abstract: Data mining is the process of analyzing data from various panoramas and epitomizes it into valuable facts. Clustering is a practical unsupervised data mining task that segregates an input data set into a required count of subgroups so that members will have high similarity and the member of different groups have large differences. K-means is a usually used for partitioning based clustering technique that identifies the user designated clusters (k), symbolized 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 approaches 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 the exhaustive exploration to discover the result to a clustering problem. The techniques typically starts with a set of randomly generated individuals called the population and procreate successive, latest generations of the population by genetic operations such as natural selection, crossover, and mutation. Each one chromosome of the population represents K no. of centroids. Stages of genetic algorithm are frequently employed for a no. of generations to search for suitable cluster centers in the trait such that an equivalence metric of the resultant clusters is optimized. The source data of K-means clustering and genetic algorithm are collated in this paper on the basis of their functioning principle, advantage and disadvantage with proper example.

Keywords: Data mining, Clustering, K-means, Genetic algorithm

I. INTRODUCTION

Data Mining is a knowledge mining process. It is an interdisciplinary regime of computer science [1]. It is the enumerated procedure of discovering patterns in bulky data sets jumbled with artificial intelligence, machine learning, statistics and database systems. The tremendous evolution of scientific databases levied a plenty of venture prior to the research to extract useful information from them using conventional data base approach. Consequently, productive mining methods are important to discover the implicit information from massive databases [2]. Clustering is one of the outstanding 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 suitable limit of subgroups such that the objects of the same subgroup will be similar or associated with each other and dissimilar or separated from the entities in other groups. Highly eligible clustering technique yields superior quality clusters with enormous and feeble intra-cluster likeliness. The excellence of a clustering outcome is determined by the likeliness and its execution and also by its capability to explore few or entire of the concealed behavior [3]. K-means is generally a partitioning based clustering method that identifies the user stipulated unit of clusters (k), which are represented by their centroids, thereby reducing the square error function. The clustering technique focuses at optimizing the cost thereby reduces the dissimilarity among the objects within each cluster, while maximizing the dissimilarity of non-identical clusters.

Genetic Algorithm (GA) parallel search method, that hunts for a universal approachable result to the clustering unease via the operation of the postulates of natural selection [4]. The algorithm conventionally commences from the set of random chosen solution called the population and builds successive, fresh propagation of the population by using genetic manipulators like natural selection, crossover, and mutation. Natural selection is discharged on the fitness of an individual. An individual, the more its fitness, has the opportunity to persist in the successive genesis. Crossover is performed by swapping the components by 1-point or 2-point crossover rule and mutation is intended to alter the string either 0 to 1 or 1 to 0 by the user-specified arbitrary position. The perception fundamental procedure is that every reborn population is superior to the earlier one. Usually, the result is illustrated by utilizing the specific portion of the strings, especially, the binary strings, but optional encodings are being evolved [6]. The principal benefit of genetic algorithms is that the fitness function is modified to innovate the performance of the algorithm.

II. LITERATURE SURVEY

Agustin Blas et al.[7] described the performance of the grouping Genetic algorithm in clustering, 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 difficult situation. 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.[8] discussed the performance of the Genetic algorithm based attribute clustering process were improved based on the grouping Genetic algorithm. The chromosome representation, Genetic operations, and fitness function defined in grouping Genetic algorithm for solving the clustering problem. 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. [9] proposed a new mode to symbolize the clustering problems casting on the likeliness between the objects and the method to combine the interior indication for ranking algorithms built on the execution of the problem. The experimental results indicated the viability of meta learning systems for an undefined mode 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. [12] enhanced reduce the local minima using evolutionary and population based methods like Genetic algorithm and teaching learning based optimization. The data sets iris and wine 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. [13] proclaimed a brief study of Genetic algorithm based clustering. Rajashree Dash et al. [10] discussed on comparative analysis of K-means and Genetic algorithm based on clustering. Arun Prabha et al.[14] with respect to the idea were 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. [15] depicted an enhanced K-means Genetic algorithm for optimal clustering. The author overcomes the disadvantage of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the technique produced more than the 90% accuracy for real life dataset. The author also adopted a neighborhood knowledge strategy for optimizing multi objective troubles. This algorithm used k means Genetic algorithm to find the smallness of the clusters. It is noted that the algorithm could produce minimum index value for the maximum datasets.

III. METHODOLOGY

3.1 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 by sketching the statistical features like means and standard deviations, perceivable evaluation it using chart and inspect for probable meaningful links among variables such as values that often occur together (Edelstein 1998).

Figure 3.1 An overview of steps that compose KDD process

3.2 K-Means Algorithm

K-means [5] clustering is a partitioning technique to classify/group the items into k groups (where k is user specified number of clusters). The group is accomplished by reducing the sum of squared distances (Euclidean distances) among the items and the matching centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of standardized density". While K-means is elementary and utilized for a vast category of data types, it is moderately reactive to the starting level of cluster centers. Two elementary techniques to cluster center induction i.e. either to select the starting values randomly, or to select the initial k samples of the data points. As an option, unique series of original values are selected from the data points and the series, that is nearer to the optimal, is opted [11].

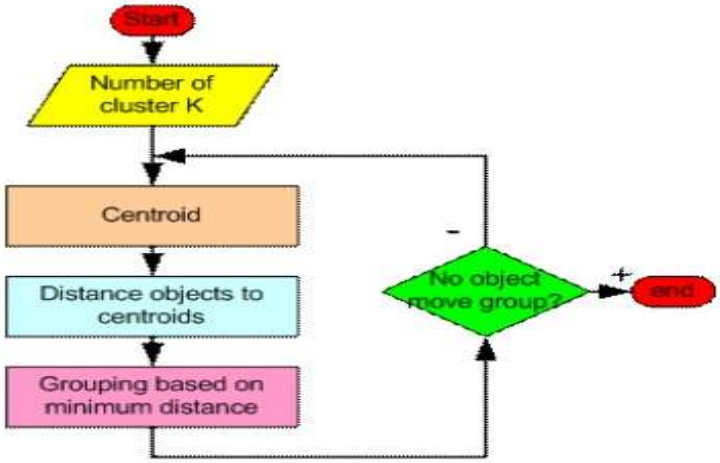

Figure : 3.2 K-Means Clustering Works

3.3 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 build on the natural selection and genetics.

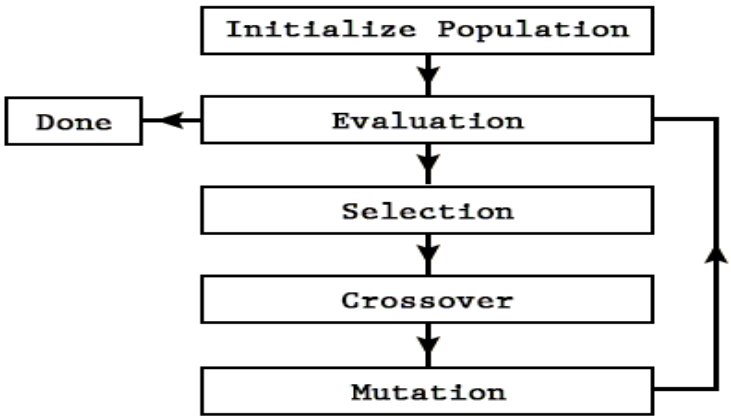

Figure: 3.3 Genetic Algorithms Overview

3.4 Proposed Algorithms

3.4.1 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= $\{x_1, x_2 ... x_n\}$.

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

3.4.2 Genetic Algorithms

Genetic algorithm facts are as follows:

- Heuristic Search Algorithms Method based on the progression of natural selection and genetics.
- Provide efficient, effective algorithms for optimization
- Beneficial when probing area is extremely high or more complicated for analysis.

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. Mushroom, Irish, Soybeans dataset has been used with

119 items for experimentation. A set of association rules are determined by using K-Means and Genetic Algorithm. By analyzing the data, and providing various support and confidence values, execution time, can obtain different number of rules. During analysis it is found that Genetic is much faster for huge number of transactions as compared 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, two types of techniques are used to find out the support, confidence, memory space, execution time accuracy of mushroom, Irish, soybeans data set. High accuracy achieved through K-Means technique compare than GA .

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 time taken by 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

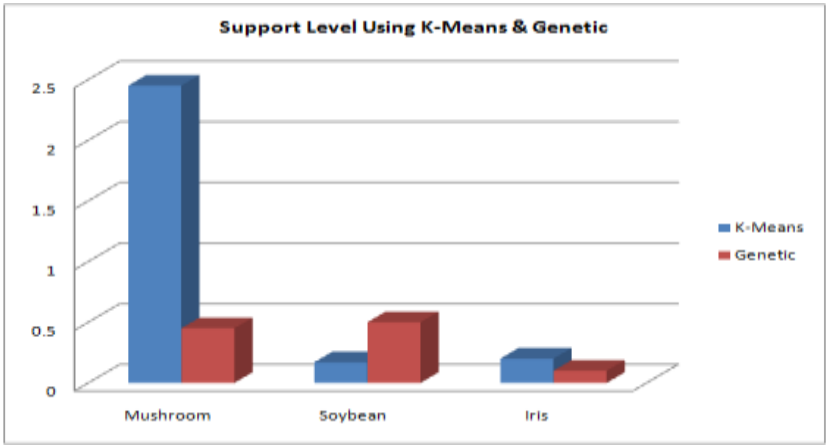

Chart 4.2(c) Graph representing Minimum Support

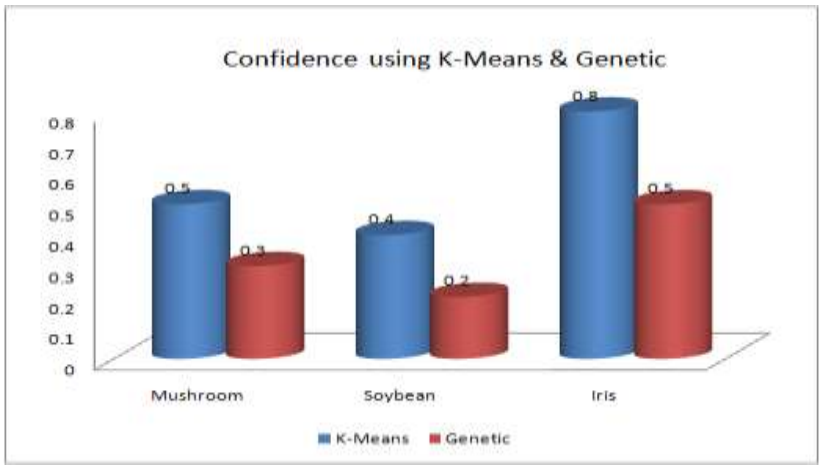

Figure 4.2(d) Graph representing Confidence

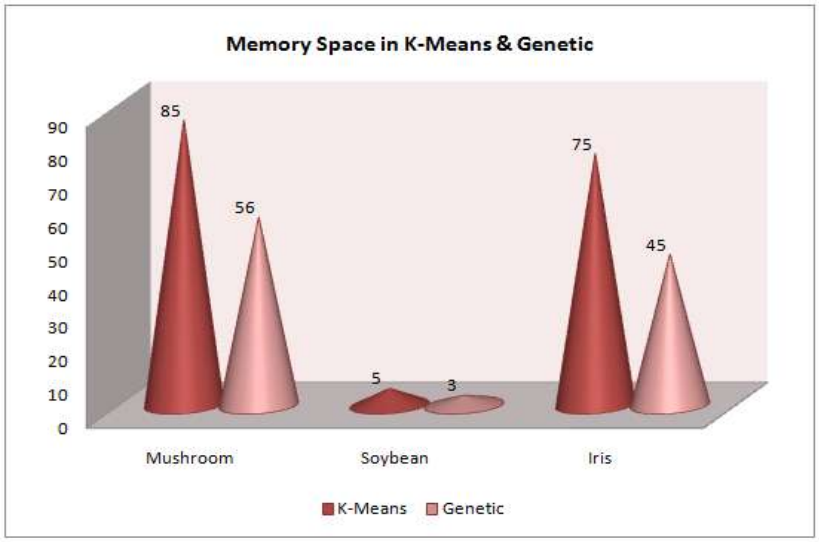

Chart 4.2(e) Graph representing memory Space

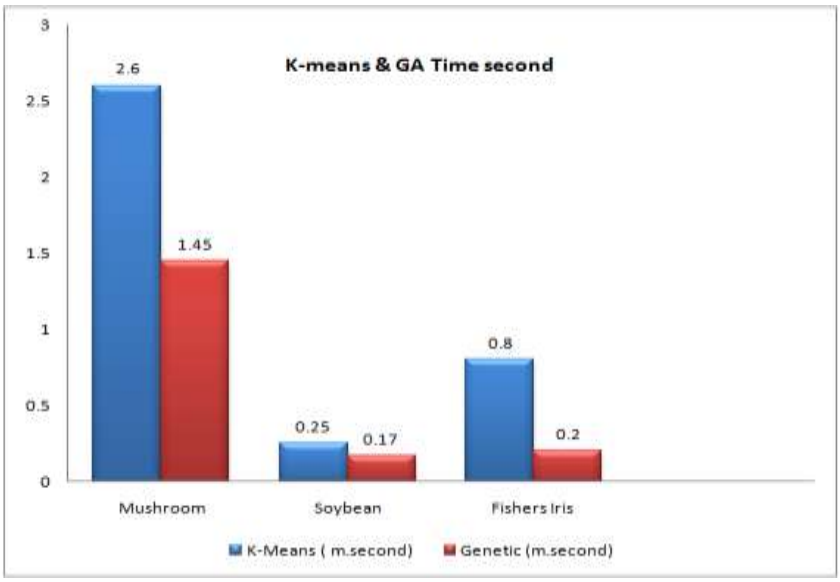

Chart 4.2(f) Graph representing Various Dataset measured in Seconds

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

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

V. CONCLUSION

In this comparative study K-means & GA techniques are used to find out the support, confidence, memory space and time in seconds of Mushroom, Soyabean and Fishers Iris data. High accuracy achieved through GA technique when compared to K-Means algorithm.. 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 limiting conditions, a GA-based clustering technique is expected to provide an optimal clustering, 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]. 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.

[6]. 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.

[7]. 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.

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

[9]. 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.

[10]. 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.

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

[12]. 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.

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

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

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

.

AUTHORS

Dr. D. Ashok Kumar did his Master Degree In Mathematics And Computer Applications In 1995 and Completed Ph.D., On Intelligent Partitional Clustering Algorithm's in 2008, From Gandhigram Rural Institute – Deemed University, Gandhigram, Tamilnadu, INDIA. He Is Currently Working as Associate Professor and Head in the Department of Computer Science And Applications, Government Arts College, Trichirappalli- 620 022, Tamilnadu, INDIA. His research interest includes Pattern Recognition And Data Mining by various Soft Computing approaches Viz., Neural Networks, Genetic Algorithms, Fuzzy Logic, Rough Set, etc., Cell: +91- 9443654052.

.

T. A. Usha completed her M. C. A. degree , from Bharathidasan University, Tiruchirappalli, Tamil Nadu. Currently doing research in Frequent Itemset Mining Techniques and Genetic Algorithm at Bharathidasan University, Tiruchirappalli, Tamil Nadu. She is currently working as Assistant Professor in the Department of Computer Science and Applications, Government Arts College, Trichirappalli- 620 022, Tamilnadu, INDIA. Cell: +91-9944429036.

C. Sivaranjani is a Full-Time M.Phil scholar in the Department of Computer Science and Applications, Government Arts College, Trichirappalli-620022,Tamilnadu, INDIA. She completed her M.Sc. degree, from Bharathidasan University, Tiruchirappalli, Tamil Nadu. She secured University Rank,while studying B,Sc. Computer Science at Holy Cross College (Autonomous), Tiruchirappalli, Tamil Nadu. Her area of interest includes Data Mining by using Genetic Algorithms. Cell: +91-8903648512.