An Enhanced K-Means Genetic Algorithms for Optimal Clustering

M.Anusha
Department of Computer Science
Bishop Heber College
Trichy-17, Tamilnadu, INDIA.
anusha260505@gmail.com

J.G.R.Sathiaseelan
Department of Computer Science
Bishop Heber College
Trichy-17, Tamilnadu, INDIA.
jgrsathiaseelan@gmail.com

***Abstract*—K-means algorithm is sensitive to the initial cluster centers and clustering results diverge with different initial input which in turn falls into local optimum. Genetic Algorithms are randomized searching technique which provides a better optimal solution for fitness function of an optimization problem. This paper proposes an enhanced K-means Genetic Algorithm for optimal clustering of data (EKMG). The aim is to maximize the compactness the clusters with large separation between at least two clusters. The superiority of EKMG is compared with grouping genetic algorithm (GGA) by using real-life dataset. The experiment shows that EKMG reaches better optimal solution with high accuracy.**

Keywords—K-Means, Genetic Algorithm, Silhouette index, Euclidean distance, clustering.

I. INTRODUCTION

Clustering of data is an unsupervised classification technique, where there is no prior knowledge of data set is available. In data clustering, patterns of the same group are similar and patterns in different group are dissimilar. A good clustering must be more compact and highly diverse from other [1].At present, this technology is widely used in information technology, biology, remote sensing and soon. It is better to study the significance of clustering which distinguishes the differences and recognizing the similarities. K-Means methodology is one of the basic methods of clustering technique which is mainly considered for its simplicity, speed, and high convergence and efficiency while dealing with high dimensional data [2].

Genetic algorithms (GA) [3-4] are randomized search and optimization technique. It works on the basic principle of evolution and natural genetics. GA performs search in huge data and provide a near optimal result for an optimization problem. The solution purely depends on its objective or a fitness function. It virtually simulates the selection, crossover and mutation phases naturally and retains the set of candidate solutions in each iteration. The better individuals are selected through genetic operators. By using genetic operators a new generation can be obtained. It is required to repeat the process until the termination condition is reached. At present, evolutionary algorithms (EAs) are widely applied for clustering problems. Since EA can hold high dimensional data also applied to large problems, it is considered as an efficient method to handle many constraint with a very trivial change. In recent research works, there are many direct application of EA to clustering problems are found. Chang et al. [16] proposed an evolutionary algorithm for clustering. In this work, objective function is constructed through similarity based constraint function that simulates the messaging between objects and optimal clusters. The performance of this method is tested with sample UCI real-life and synthetic data sets [5]. Liu et al. [6] proposed a real coding evolutionary algorithm with special division namely, absorption mutation operator and a fitness function based on the Davis-Bouldin index. Similar such works dealt with evolutionary clustering are Xia et al. [7], Zahraie et al. [8]. There are also several evolutionary approach involved in clustering are particle swarm optimization [9], ant colony optimization [10], bee colony optimization [11] and multi-objective optimization [12]. In spite of massive application of evolutionary technique, we propose an Enhanced K-Means Genetic algorithm for clustering compact huge data to overcome local optima. In this paper, we present a contribution to clarify the performance of Enhanced K-Means Genetic Algorithm for clustering problem. The proposed work includes new encoding of genetic operators and the performance of this algorithm is studied with different fitness function. The experiments on UCI real-life data sets is used to complete our study on Enhanced K-Means Genetic algorithm.

The rest of the paper is designed as follows: section 2 summarizes methodology of K-Means and Genetic algorithm. Section 3 presents Enhanced Hybrid K-Means Genetic Algorithm proposed in this paper for clustering problems. Section 4 contains experimental part of this paper, where the performance of the proposed algorithm is evaluated. Section 5 closes this paper by giving remarks and conclusions.

II. METHODOLOGY

Clustering involves partitioning of data sets into similar group. In general, clustering algorithms are categorized into two types: hard clustering problems and soft clustering problems. In hard clustering each data point belong to one and only cluster where as in soft clustering each data is allowed to have link with more than one cluster [13-14].

A. K-Means Clustering

K-Means clustering is the fast and simple method for its straight forward method with smaller number of iterations. This algorithm divides the data set into K disjoint subsets. The cluster requirement is estimated based on the user's choice.

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

The computer randomly select and assign the object to one of the clusters (k). The distance between each object and the center of each cluster is calculated which results an optimal cluster solution. The objects inside the particular cluster are close to each other.

B. Genetic Algorithm

The Genetic algorithm is used to cluster huge data for its ease implementation and accessibility. The searching capability of GA is applied to define the cluster centers 'k' for the unlabeled data 'n'. Clustering fitness is calculated to sum the objects with the cluster centers using Euclidean distance. Mathematically, a clustering metric M for 'k' number of clusters is stated as

$$M(c_1, c_2, \ldots, c_k) = \sum_{i=1}^{k} \sum_{x_j \in c_j} \|x_j - CC_j\| \quad (1)$$

The aim of GA is to search for appropriate cluster centers $CC_1, CC_2 \ldots, CC_k$ such that clustering metric is minimized.

III. Proposed Algorithm: Ehhanced K-Means Genetic Algorithm

The disadvantage of K-Means algorithm can be treated as a main research problem. Hence the result of the search path often confined in a local minimum, if the initial setting in not course that will lead to the global solution. The proposed EKMG algorithm attempt to overcome this problem by appending the K-Means with a searching technique. The method is motivated by the concepts of Genetic Algorithm, which is an optimal search algorithm. In this crux, the EKMG algorithm preserves the basic framework of K-Means and adds population on the top of the trail. Based on the fitness values a new population is generated. The value of the centroid is considered as the fitness for that particular group. Arithmetically, we have calculated centroid as

$$newCentroid(j, p) = newCentroid(j,p) + data(i,p) \quad (2)$$

where, j and p are the data points of the class i. The distance is calculated as

$$sum = sum + (data(j,p) - centroid(i,p))\text{^}2; \quad (3)$$

The data is defined as

$$data = rand(size, coordinates); \quad (4)$$

where rand() takes the value randomly from the data.

size specifies the coordinates of data and the coordinates are columns corresponds to the targeted data. The cluster can be formed from the HKMG algorithm is as follows

$$[clust\_group\ Fun] = ga\ (cluster, k, options) \quad (5)$$

where 'cluster' is the cluster-objective function using K-Means with *k* cluster having the options of GA like population size, Crossover function, mutation operator and soon. The termination state can be achieved, if the targeted results are obtained or exceeds the generation or if the fitness value is less than the stall generation limit. It is because the centroids are tested in genetically K-Means environment. The results are evaluated after the completion of single HKMG iteration. Hence, the best results obtained while iteration is considered as optimal and a final result.

A. Algorithm Design for Proposed Work

1. Initialization
   - Input Number of Cluster *k*, Population *N*, Number of Generations *G*, Crossover function, Mutation Function.
   - Choose the initial cluster center based on fitness.
2. Evaluation of clusters
   - Obtain new cluster centers by K-means Algorithm.
   - Check whether the selected object is the main object
   - Cluster the object according to new cluster centers and calculate the fitness based on the (2);
   - Assign all closely linked objects as a cluster.
   - Keep the best individuals as the next generation.
3. Generation of new individuals by genetic operations
   - Selection based on the fitness of newly generated population.
   - Single-point crossover on the new population.
   - Form-mutation for the selected population.
4. Stopping Criteria
   - When number of iterations exceeds generations, stop the iteration and select the best fitness value as the result of algorithm. Otherwise, go to step 2 for iteration again.

IV. Experiments and Results

In this section, a comparative experimental evaluation of the proposed algorithm is presented.

A. Datasets

The proposed work is evaluated with one high dimensional data set and a low dimensional data set which is given in Table I. the data sets are obtained from public warehouses. Its performance is compared against the Grouping Genetic Algorithm (GGA) [15]. The detailed description of this data sets is found in UCI repository. The size of the datasets are M, number of attributes are defined as P and K_c are the real number of clusters.

TABLE I. Main Features of Datasets

Data Sets	*M*	*P*	K_c
Iris	150	4	3
Wine	178	13	3

B. *Comparison of Cluster Accuracy*

In order to evaluate a clustering method, it is necessary to define a measure of agreement between two partitions of same data sets. The first experiment is carried out with the Iris problem. This problem is considered as main data mining clustering problem. It is formed by 3 classes having 4 features like length and width of sepal and petal of the flower, in centimeters. The Wine problem from UCI repository is considered as second experimental. Wine is formed by 3 cluster classes with 178 data objects.

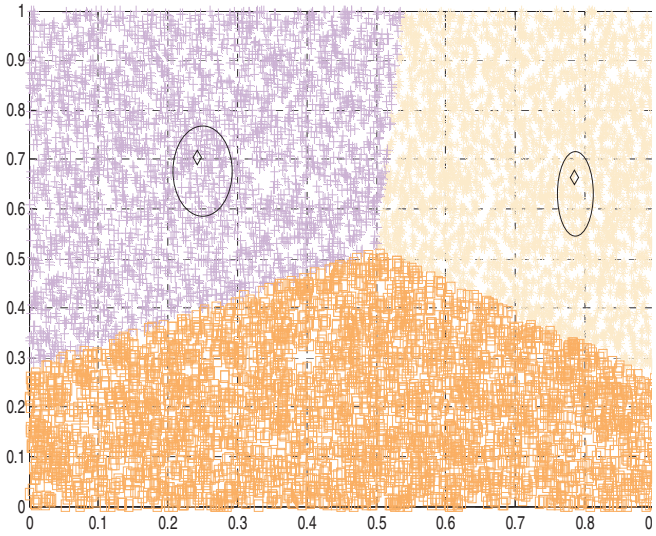

Fig.1. Best clustering obtained with the proposed EKMG with Silhouette index for Iris problem considered for clustering.

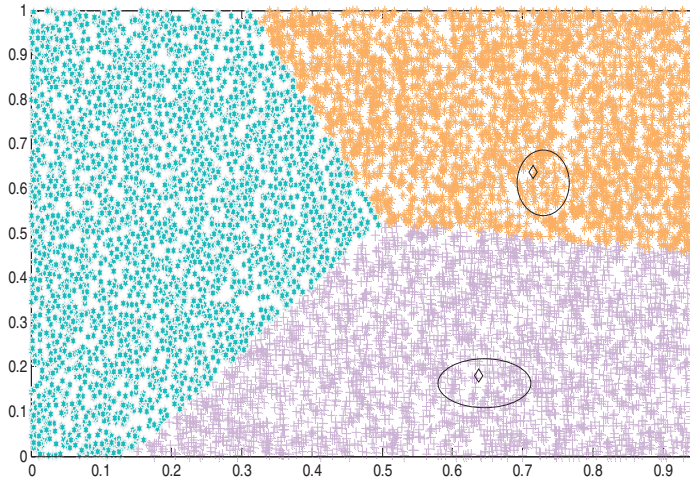

Fig.2. Best clustering obtained with the proposed EKMG with Silhouette index for Wine problem considered for clustering.

Table II. shows the results obtained from GGA and the proposed algorithm. The quality of the cluster is evaluated using silhouette coefficient (S) and the objective of a good cluster partition is to maximize S.

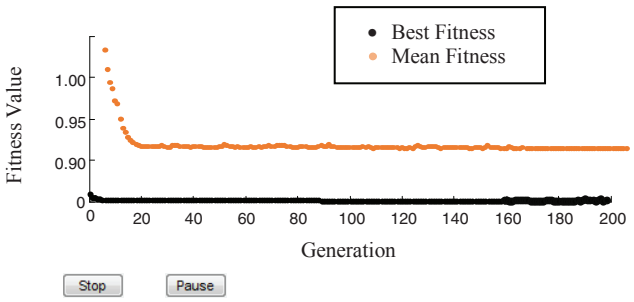

Fig. 3. EHKMG plots of algorithm on Wine problem

TABLE II. THE RESULT OF CLUSTERING ACCURACIES OF GGM,EHKMG

Data Sets	*GGM*	*EHKMG*
	Silhouette S(i)	
Iris	0.8995	1.0204
Wine	0.72220	1.0000

From the above result, it is certain that the clustering accuracy of EKMG is better than GGM. The clustering accuracy values of GGM is smaller than proposed EKMG. This is because of inadequate cluster numbers identification. Moreover, all of the index values of EKMG for two real-life data set is 1. Hence, we can conclude that EKMG is more efficient than GGM for clustering problems. The efficiency of EKMG is obtained by searching the cluster centers based on

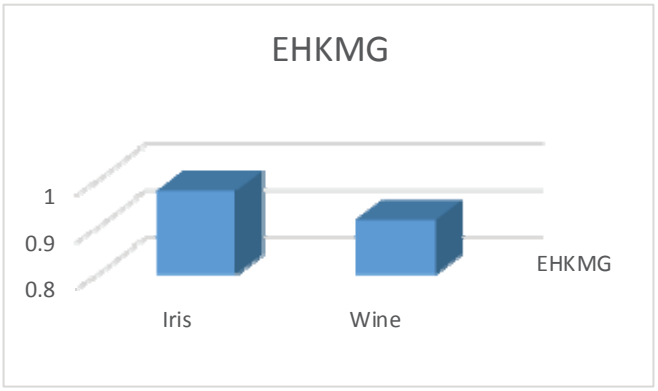

Fig.5. EHKMG chart for cluster group-ratio identification

the population fitness. Hence, the algorithm can easily accomplish the complex data space of the population.

C. *Identification of Compact Cluster*

To analyze the searching capability of proposed algorithm, we compared EKMA with the GGA. The GGA algorithm modifies the genetic operators of traditional GAs to obtain a compact algorithm to determine the number of clusters. In this

experiment on identification of compact cluster is devised by setting parameters of EKMG as follows:

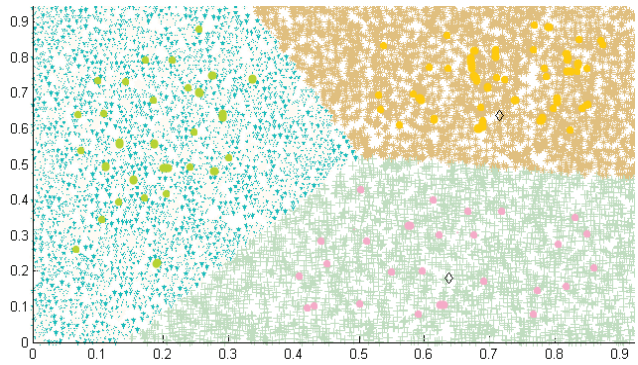

Fig.6. Cluster group identification of EHKMG on Iris

Population size N=100
Generation G_i =200
Crossover =singlepoint crossover
Mutation =form mutation

The experiments are conducted on Matlab 7.0 software package. The results of cluster group identification of real-life problem is shown in Fig 6. EKMG uses entire information in the dataset. This shows the presentation of global search in data sets for a good cluster separation. Size of the cluster depends on the population size. Table III. shows the value of cluster ratio (cl_r) obtained by EKMG. It is inferred that cl_r

TABLE III. THE CLUSTER GROUP –RATIO IDENTIFICATION OF EKMG

Data Sets	*EHKMG(cl_r)*
Iris	0.97607
Wine	0.91550

values of EKMA is more than 90% for real-life problem. The results of cluster group-ratio identification is comparatively high for Iris than Wine problem.

D. Time Complexity of EKMG

Time complexity of an algorithm is considered as a major issue for a clustering problem. The Iris problem is taken into consideration for estimating the time complexity of the proposed algorithm. It is inferred that when the population generation is set to 200 and the population size is incremented by 10 automatically. Hence, the increase in population size increases the processing time of generations. The experiments were implemented on compaq laptop with Intel Core processor, 16 GB memory and windows 7. Therefore, the total time complexity of EKMG is O ($NG^{2)}$.

V. CONCLUSION

Clustering is a common way of identifying the complex group data. Generally, the clustering speed of K-Means is fast it often suffers from local optimum for a huge data. Hence, the results need to be optimized. By applying genetic algorithm for clustering the small data, there is possibility to reach global optimal solution. However, it is not suitable for large data sets. The proposed enhanced k-means genetic algorithm for clustering overcomes the above drawback. This algorithm is well suitable for low volume data and a large-complex data. However, the time complexity of the algorithm is relatively high. There is a need of improvement in cluster number identification.

The experimental results on two real-life datasets confirmed that EKMG has better performance than GGA in clustering problem. The proposed method can able to find more compact clusters. The clustering accuracy of the EKMG is good when compared with GGA. Therefore, the future work would be to study in improving the EKMG to reduce the diversity on the huge datasets.

REFERENCES

[1] Satyasai Jagannath Nandaa, , , Ganapati Pandab, "A survey on nature inspired metaheuristic algorithms for partitional clustering", Swarm and Evolutionary Computation, Volume 16, 2014, Pages 1–18.

[2] M.C. Naldi, , R.J.G.B. Campello , E.R. Hruschka , A.C.P.L.F. Carvalho, "Efficiency issues of evolutionary k-means", Applied Soft Computing, Volume 11, Issue 2, 2011, Pages 1938–1952.

[3] Chih-Fong Tsaia, , , William Eberleb, Chi-Yuan Chua, "Genetic algorithms in feature and instance selection", Knowledge-Based Systems, Volume 39, 2013, Pages 240–247.

[4] András Királya, , , Asta Laihob, János Abonyia, Attila Gyeneseib, "Novel techniques and an efficient algorithm for closed pattern mining", Expert Systems with Applications, Volume 41, Issue 11, 2014, Pages 5105–511.

[5] Núria Macià, , Ester Bernadó-Mansilla, "Towards UCI+: A mindful repository design", Information Sciences, Volume 261, 2014, Pages 237–262.

[6] Liu, Y., Wu, X., and Shen, Y , "Automatic clustering using genetic algorithms", Applied Mathematics and Computation, Volume 218, Issue 4, 2011,Pages 1267–1279.

[7] Xiao, J., Yan, Y., Zhang, J., and Tang, Y. A, " quantum-inspired genetic algorithmfor K-means clustering.", Expert Systems with Applications, Volume 37,Issue 7, 2010. Pages 4966–4973.

[8] Zahraie, B., and Roozbahani, A, "SST clustering for winter precipitationprediction in southeast of Iran: Comparison between modified K-means and genetic algorithm-based clustering methods." Expert Systems with Applications, Volume 38, Issue 5, 2011. Pages 5919–5929.

[9] Shafiq Alama,Gillian Dobbiea, Yun Sing Koha, Patricia Riddlea and Saeed Ur Rehmanb, "Research on particle swarm optimization based clustering: A systematic review of literature and techniques", Swarm and Evolutionary Computation, Volume 17, 2014, Pages 1–13.

[10] Md. Monirul Kabira, Md. Shahjahanb and Kazuyuki Murasea, "A new hybrid ant colony optimization algorithm for feature selection", Expert Systems with Applications, Volume 39, Issue 3, 15 2012, Pages 3747–3763.

[11] Oleynik, A., Subbotin, S. , "Bee colony optimization for clustering", Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), 2010 International Conference ,2010, Page(s): 286.

[12] Anirban Mukhopadhyay, Sanghamitra Bandyopadhyay, "A survey of multiobjective evolutionary algorithms for data minig: Part-II", IEEE Trans. Evolut. Comput, 2014.

[13] Francisco de A.T. ,de Carvalhoa, Yves Lechevallierb and Filipe M. de Meloa, "Partitioning hard clustering algorithms based on multiple dissimilarity matrices", Pattern Recognition, Volume 45, Issue 1, 2012, Pages 447–464.

[14] Lin Zhu ; Inst. of Image Process. & Pattern Recognition, Shanghai Jiao Tong Univ., Shanghai, China ; Longbing Cao ; Jie Yang, "Soft subspace clustering with competitive agglomeration", Fuzzy Systems (FUZZ), 2011 IEEE International Conference,DOI: 10.1109/FUZZY.2011.6007424 ,Publication 2011, Pages 691 - 698.

[15] L.E. Agustı´n-Blas , S. Salcedo-Sanz , S. Jimenez-Fernandez , L. Carro-Calvo, J. Del Ser and J.A. Portilla-Figueras , "A new grouping genetic algorithm for clustering problems", Expert Systems with Applications, Expert Systems with Applications, Volume 39 ,2012, Pages 9695–9703.

[16] Chang, D., Zhao, Y., Zheng, C. and Zhang, XA, " genetic clustering algorithm using a message-based similarity measure", Expert Systems with Applications,Volume 39,Issue 2, 2012, Pages 2194–2202.