An Unsupervised Classification of Cancer Genes Using Genetic Algorithm

Helan Cynthiya Y., Anusha M., Dr. J. G. R. Sathiaseelan

Abstract— **K-means algorithm is a standard unsupervised clustering technique that helps in successful termination of problem in relatively efficient manner. Though K-means is simple and easy to implement, it fails to find the most optimal configuration as the initialization of clusters at the beginning is difficult and sensitive to cluster centers. To improve the efficiency of K-means, Evolutionary Algorithm, like Genetic Algorithm is used for solving local optima obtained from the former. The algorithm EKMGA is proposed in this paper, which finds the fittest value of the chromosome at the minimum number of generation with K-means selection of clusters. The algorithm EKMGA is tested using two different cancer datasets that were taken from UCI repository. The implementation is done using Java code and Weka tool.**

***Index Terms*— Clustering; K-means Algorithm; Genetic algorithm; Enhanced K-means Genetic Algorithm;**

I. INTRODUCTION

Clustering is an unsupervised technique that groups the data of similar objects with minimum cluster distance into a cluster by eliminating the inappropriate data objects [1]. Different types of clusters are well- separated cluster, contiguous cluster, shared property or conceptual cluster, center-based cluster and density based cluster. Clustering discovers the close groups within the data and helps in finding the knowledge discovery in engineering and scientific domains such as psychology, medicine, remote sensing, etc.. Hierarchical clustering, Partition based clustering and grid based clustering are the types of clustering algorithm. Clustering is of two types namely, hard clustering and soft clustering. Each element in a sample belongs to exactly one cluster comes under hard clustering. Each element belongs to each cluster in a population refers to soft clustering [2].

K-means algorithm is an unsupervised and simple partition - based clustering method that aims to group *n* observations into k clusters with nearest mean. *k* is the number of clusters in K-means which is iterative in nature. The time complexity of K-means is given by *l*k*n*, where *l* is the number of iterations, *k* is the number of clusters and *n* is the number of observations. Typically, K- means algorithm converges in small number of iterations [3].

Manuscript received July, 2015.

***Helan Cynthiya Y.**, M.Phil., Research Scholar,Department of Computer Science, Bishop Heber College,. Tiruchirappalli,Tamilnadu, India.*

***Anusha M**, Ph.D., Research Scholar,Deparment of Computer Science, Bishop Heber College,Tiruchirappalli, Tamil Nadu,India.*

***Dr.J.G.R.Sathiaseelan M.Sc., Ph.D.,**,Research Advisor, Head-Department of Computer Science,Tiruchirappalli, Tamilnadu,India.*

Although K-means is simple and fast unsupervised clustering algorithm, it has some advantages and disadvantages. The advantages of K-means algorithm are procedure always terminates successfully, relatively efficient. Some disadvantages of using K- means algorithm are failed to find the most optimal configuration, sensitive to the initial randomly selected cluster centers, applicable only when mean is defined, need to specify number of clusters K which is difficult [4].

Genetic Algorithms is a heuristic search Algorithms which is based on natural selection and genetics, an evolutionary ideas inspired by Darwin's evolution theory [5]. GA simulates the survival of the fittest among individuals and depends on the genetic structure, behavior of respective chromosomes within the population of individuals. The basics of GA could be stated as individuals in a population compete for resources and mates, the individuals with most successful probability in each 'competition' would produce more offspring than the individuals that perform poorly, Genes from 'good' individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent, thus each successive generation will become more suited to their environment. Population of individuals is maintained within the search space. Each individual is coded as a finite length vector component represented as binary {0,1}. The set of genes forms a chromosomes and a fitness score is assigned to each individuals to compete. The individual with optimal fitness score will be taken into account for mating of parents to produce better offspring. New arrival of offspring replaces the individuals with least fitness score [6]. This paves the way for the best solution in the successive generation.

Population repository maintains the maximum number of chromosomes along with their fitness values. Parents with better fitness value are allowed for mating to produce offspring that inherit the characteristics from their parent. Representations of chromosomes [7], size of population, fitness function, selection [8], crossover [9], Mutation [10]. This paper is organized as follows: Section II discusses a survey on clustering based Genetic Algorithm, Section III describes the proposed algorithm EKMGA, Section IV depicts the implementation of the algorithm and their results and discussions and Section V draws a conclusion along with the future enhancement.

II. LITERATURE REVIEW

Sonia et al. [11] anticipated a hybrid method called Genetic K- means algorithm which prevents from converging towards local minima. The algorithm is tested

over 2D and 3D data for 500 points with initial population of 500 chromosomes. The execution time for 2D and 3D data has been obtained in millisecond along with their number of generations. The time complexity and execution expectation has also been tested over exhaustive set of data of different dimensions.

Fang et al. [12] depicted a hybrid algorithm of the weighted k- means algorithm and a genetic algorithm known as Genetic Weighted K- Means Algorithm (GWKMA). The algorithm was experimented on one synthetic and two real-life gene expression datasets. The obtained results proved that the performance of GWKMA is better than the simple K-means in terms of the cluster quality and the clustering sensitivity to initial partitions and it overcomes the disadvantages of the K –means and Weighted K-means. In addition, the algorithm GWKMA was found to be generic and has the application on clustering large scale biological data such as gene expression data peptide mass spectral data.

Varsha et al. [13] investigated the use of Genetic algorithms to determine the best initialization and optimization of clusters. For demonstrating the effectiveness of Genetic Algorithm, two artificial datasets with the number of clusters ranging from 2 to 2000 has been considered. Hence it is assured that the performance of Genetic Algorithm was significantly superior to that of the K-means algorithm.

Rajashree et al. [14] compared the analysis of K- means and Genetic Algorithm on the basis of their working principle, advantages and disadvantages. The comparison of these two algorithms has been done on the same datasets with 6 data objects and 2 variables. The experimental results stated that the K-means converges to local optima and Genetic Algorithm performs global search approaches with implicit parallelism. The authors have concluded that GA based clustering provides better optimum solution compared to that of K-means algorithm.

Dharmendra et al. [15] proposed a Fast Genetic K- means algorithm that works well on the datasets with mixed numeric and categorical features. The datasets taken into consideration are iris, wine, heart disease, etc.. The result of Heart disease dataset is tabulated along with their performance analysis. The dataset contains 303 instances with two classes namely normal patient (164) and heart patients (139). The FGKA algorithm converges very fast compared to that of GKMODE and IGKA.

Bouhmala et al. [16] combined Genetic Algorithm and K-Means to improve the quality of clusters formed and speed up their search process. The performance of GAKM is tested over the datasets such as iris, glass, etc., and that has been taken from Machine learning repository. The experimental results have proved that GAKM converges faster while comparing to standard Genetic Algorithm. Though this algorithm failed to capture the best quality of clusters, it is unsuitable for the maximizing both homogeneity and heterogeneity within same clusters and with different clusters respectively.

Rouhollah et al. [17] projected a model called GA clustering for improving K-means algorithm. The algorithm has been performed on well known datasets namely iris, crude oil. The experimental results provide clustering standard μ, the lower the value of μ gives the better cluster formation of data compared to traditional K-means clustering algorithm.

Kailash et al. [18] presented a partition based Genetic Algorithm Initialized K-means (PGAIK) in order to improve the performance of traditional clustering technique. The PGAIK algorithm has been tested on the remote sensing image of GJU and has been analyzed that PGAIK has yielded more compactness than the Genetic Algorithm Initialized K-means (GAIK).

A vast survey on clustering based evolutionary algorithms in gene pattern mining was done by Helan et al [19]. The performance analysis of evolutionary algorithm GA is tabulated using four different medical datasets namely, colon and leukemia cancer datasets, brain tumor and lung tumor datasets. The percentage rate for colon cancer using GA, BFSSGA,GA/AIS are 98.20%, 87.54% and 87.70% respectively. Similarly for leukemia cancer datasets are 100%, 87.47% and 98.33% and has been concluded that evolutionary algorithms like GA produces near optimum solution compared to that of traditional algorithms.

Anusha et al. [20] depicted an enhanced K-means Genetic Algorithm for optimal clustering. The author overcomes the drawback of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the algorithm produced more than 90% accuracy for real life datasets.

III. PROPOSED WORK

K-means is an iterative algorithm which configures cluster centers randomly. Each pattern is assigned to the cluster whose cluster center is closest to the pattern among all the cluster centers. Genetic Algorithm is a probabilistic search algorithm that iteratively transforms a set of mathematical objects into a set of new offspring with their respective fitness value through Darwinian principle of natural selection and genetic operators such as crossover and mutation. The Enhanced K-Means Genetic Algorithm (EKMGA) assigns the k value for the chromosomes and then Genetic Operators like selection, mutation and crossover for better optimal solutions.

The coding, initialization schemes and genetic operators in K- means Genetic Algorithm is specified below:

1. ***Representation***: Assigning the value for each allele in chromosome of length as {1,2,3,…..,k}.
2. ***Population Initialization***: Initial population is chosen randomly with k number of clusters.
3. ***K- Means Operator***: In order to overcome the low value of mutation, K- means operator is used. The two steps that contribute to improve this situation are:
   - Calculate the cluster center.
   - Reassign each data point to the cluster with the nearest cluster center and thus form best solution.
4. ***Selection***: Selection operator randomly select a chromosome from the previous population.

Algorithm : EKMGA

1: **Input**: Dataset X

2: **Output**: Clusters C

3: **Begin**

4: **Generate** Initial population

5: **Assign** K value for each chromosome in a population.(see Pseudo Code)

6: **Calculate** fitness for each individuals in a population

7: **While**(not convergence reached) do

a. Select individuals with better fitness

b. Perform crossover to produce offspring

c. Apply mutation if necessary

d. Replace the old individual in the population with the new generation

8: **End**

9: **End**

5. ***Mutation***: The mutation changes the allele value depending on the distances of the cluster centroids from the corresponding data point. The probability of changing an allele value to a cluster number is more if the corresponding cluster center is closer to the data point.

A general pseudo code for Genetic K-means Algorithm is as follows:

Pseudo Code

1: **initialize** each chromosome to contain k randomly chosen centroid from the datasets. /*k- number of clusters*/

2: **for** t=1 to max /* t- maximum number of iteration */

3: **for** each chromosome i

a. **assign** the data object to the cluster with the closest centroid.

b. **recalculate** k cluster centroids of chromosome i as the mean of their data objects.

4: **calculate** the fitness of chromosome i.

5: **create** the new generation of chromosomes using selection, crossover and mutation.

The algorithm EKMGA is described below with the initialization of population, k value for each chromosome in the population is assigned for computation and the Genetic operators like selection, crossover and mutation are performed. The convergence mentioned in the algorithm is the criteria at which the population tends to lose its diversity. The rest of the genetic operators like selection perform natural selection of chromosome with better fitness value, crossover perform the task of producing offspring and mutation perform to produce mutated offspring if necessary.

Firstly, the algorithm generates initial population and then it proceeds for k value assignment to each chromosome in the population. EKMGA checks for the cluster center for assigning fitness value, otherwise, it goes to for loop until it recognizes the cluster center. Then the genetic operators like selection of best chromosome based on fitness value are performed followed by crossover for producing better offspring. If necessary the mutation operator is performed. The process is terminated when the expected optimum result is obtained.

IV. Experimental Results and Discussions

The experiment is done on the two different datasets such as Breast cancer datasets and Lung cancer datasets using Java coding and Weka tool.The implementation is done in Eclipse using JAVA code. It is tested over Windows 7 with Intel ® Core™ i3 CPU M 380 @ 2.53GHz, 3.00 GB Installed memory, 64-bit OS system.

The Breast cancer datasets and Lung cancer datasets has been taken from UCI Repository. The Breast cancer dataset has the instances of 699 with 10 attributes and the lung cancer dataset has the instances of 32 with 56 attributes is tabulated in TABLE I.

TABLE I. Dataset Description

S.No	Dataset Description		
	Dataset	***Number of attributes***	***Number of Instances***
1	Breast Cancer	10	699
2	Lung Cancer	56	32

The fittest values for the selected data sets are obtained at the minimum number of generations using Java code. The results are tabulated in Fig.4.1.

The k value obtained from the K-means Genetic Algorithm is in the range of 4-6. Similarly for Wisconsin Breast cancer dataset, the value of k is 2 and the obtained value from the Genetic K means algorithm is in between the range of 3-5 and is listed in TABLE II.

TABLE II. K value for Breast Cancer using EKMGA

S.No	K value for Breast cancer using EKMGA	
	Iterations	***K Value – EKMGA***
1	2	3.46
2	1	3
3	2	5
4	3	4

The k value for Lung cancer dataset is 3 and the obtained value from the Genetic K means algorithm is in between the range of 3-5 and is listed in TABLE III.

TABLE III. K VALUE FOR LUNG CANCER USING EKMGA

S.No	K value for Breast cancer using EKMGA	
	Iterations	*K Value – EKMGA*
1	2	3.72
2	1	4
3	2	3
4	3	5

The attributes for the Breast Cancer datasets has been obtained using Genetic Search method are tumor-size, inv-nodes, node-caps, deg-malig, irradiat. There are 17 attributes of Lung cancer datasets has been selected from 57 attributes using Genetic Search method. The fittest value for Breast cancer dataset and Lung cancer dataset has also been obtained and tabulated in TABLE IV.

TABLE IV. FITTEST VALUE OF BREAST CANCER AND LUNG CANCER

Generation	Fittest Value – EKMGA	
	Breast Cancer	*Lung Cancer*
1	19	17
2	19	18
3	24	20
4	25	21
5	26	22

The Fittest value for the datasets, Breast cancer and lung cancer is depicted using line chart. Fig. 4.1.

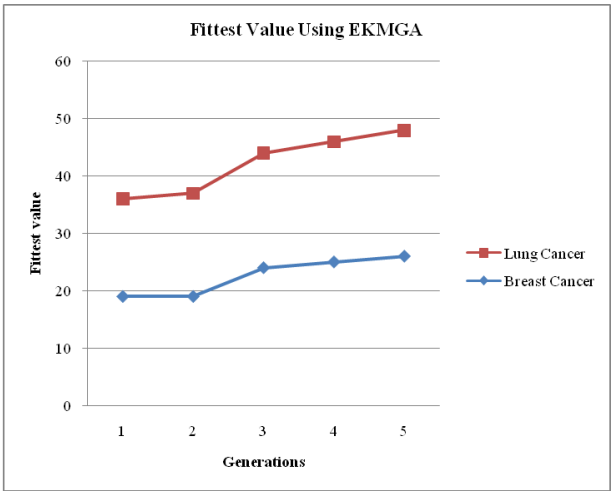

Fig. 4.1. Fittest Evaluation of Breast cancer and Lung cancer- GA

From the above analysis, it is inferred that the algorithm converges at the fittest value of 19 and 17 for breast cancer dataset and lung cancer dataset respectively at the first iteration compared to the traditional algorithms like Genetic Algorithm, K-means and Genetic K-means algorithm. Hence the proposed algorithm EKMGA performs much better in terms of finding fittest value at the minimum number of iterations with minimum number of cluster formation.

V. CONCLUSION

Clustering is an important unsupervised classification technique where a set of data objects taken in a multi-dimensional space, are grouped into clusters. In such a way, the data objects in the same cluster are similar in some patterns and objects in different clusters are dissimilar in the same patterns. K-Means is an intuitively simple and effective clustering technique. Nonetheless, K-means depends on the choice of the initial cluster centers, which normally stuck at sub-optimal solutions, whereas GA is a randomized search and optimization technique guided by the principles of evolution and natural genetics, and having a large amount of implicit parallelism. Therefore, GA provides near optimal solutions for the selected objective or fitness function of an optimization problem. Under limiting conditions, a Clustering based GA technique is expected to provide an optimal clustering which is more superior to that of K-Means algorithm with little more time complexity. The proposed algorithm EKMGA results in providing the fittest value for two selected cancer datasets with minimum number of generations. The future enhancement can be done using Multi Objective Genetic Algorithm (MOGA) and Multi Objective Particle Swarm Optimization (MOPSO) which provides the better optimal solution than that of Genetic K-means Algorithm, as they both fall under parallelism mechanism.

REFERENCES

[1] R.S. Santos, S.M.F. Malheiros, S. Cavalheiro, J.M. Parente de Oliveira, "A Data Mining system for providing analytical information on Brain tumors to public health decision makers", *Computer Methods And Programs in Biomedicine*, ISSN:0169-2607, pp. 296-282, 2013.

[2] R.Roseline, G.Jenitha, J. Henri Amirhtaraj, "Analysis and Application of Clustering Techniques in Data Mining", *International Journal of Computing Algorithm*, pp.910-912, 2014.

[3] Roohollah Etemadi, Alireza Hajieskandar, "A Novel Evolutionary Algorithm for Data Clusteringin N Dimensional space", *Indian Journal of Computer Science And Engineering*, ISSN: 0976-5166, pp. 902-908, 2012.

[4] Yogita Chauhan, Vaibhav Chaurasia, Chetan Agarwal, "A Survey of K-means and GA-KM the Hybrid Clustering Algorithm", *International Journal of Scientific & Technology Research*, ISSN: 2277-8616, pp.119-122, 2014.

[5] Gunjan Verma, Vineeth Verma, "Role and Applicationof Genetic Algorithm in Data Mining", *International Journal of Computer Applications*, ISSN: 0975-888, 2012.

[6] Richa Garg, Saurab Mittal, "Optimization by Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN:2277 128X, pp.587-589, 2014.

[7] Shaifali Aggarwal, Richa Garg and Dr. Puneet Gaswami, " A Review Paper on Different Encoding Schemes used in Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X, pp.596-600, 2014.

[8] Noraini Mohd Razali, John Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP", *Proceedings of the World Congress on Engineering*, ISBN: 978- 988-19251-4-5, ISSN: 2078-0958, 2011.

[9] Jorge Magalhaes-Mendes," A comparitive study of Crossover operators for Genetic Algorithms to solve the Job Scheduling Problem", *WSEAS Transactions on Computers*, ISSN:2224-2872, pp. 164-173, 2013.

[10] Nitasha Soni, Dr Tapas Kumar, "Study of various Mutation Operators in Genetic Algorithms", *International Journal of Computer Science and Information Technologies*, ISSN: 0975-9646, pp. 4519-4521, 2014.

[11] Sonia Sharma, ShikhaRai, "Genetic K- Means Algorithm-Implementation and Analysis", *International Journal of Recent Technology and Angineering*, ISSN: 2277-3878, 2012.

[12] Fang-Xiang Wu, "Genetic Weighted K- means algorithm for clustering large-scale gene expression data", *BioMed Central*, 2008.

[13] Varsha Singh, Prof A.K. Mishra, "A Genetic Algorithm for K-Means Clustering", *International Journal of Emerging Technologies in Complutational and Applied Sciences*, ISSN: 2279-0047, pp. 359-364, 2014.

[14] Rajashree Dash, Rashmita Dash, "Comparative analysis of K-means and Genetic Algorithm based data clustering", *International Journal of Advanced Computer and Mathematical Sciences*, ISSN: 2230-9624, pp. 257-265, 2012.

[15] Dharmendra K Roy, Lokesh K Sharma, "Genetic K- Means Clustering Algorithm for Mixed Numeric and Categorical Datasets", *International Journal of Artificial Intelligence & Applications*, pp. 23-28, 2010.

[16] N. Bouhmala, A. Viken, J. B. Lonnum, " Enhanced Genetic Algorithm with K-Means for the Clustering Problem", International Journal of Modeling and Optimization, pp. 150-154, 2015.

[17] Rouhollah Maghsoudi, Arash Ghorbannia Delavar, Somayye Hoseyny, Rahamatollah Asgari, Yaghub Heidari, "Representing the New Model for Improving K-meansClustering Algorithm based on Genetic Algorithm", *The Journal of Mathematica and Computer Science*, pp. 329-336, 2011.

[18] Kailash Chander, Dinesh Kumar, Vijay Kumar, "Enhancing Cluster Compactness using Genetic Algorithm Initialized K- means", *International Journal of Software Engineering Research & Practices*, pp. 20-24, 2011.

[19] Helan Cynthiya Y, Anusha M, J. G. R. Sathiaseelan, "Cognitive Development of Evolutionary Algorithms in Gene pattern Mining", International Journal of Computer Science and Mobile Computing, ISSN: 2320-088X, pp.366-372, 2015.

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

