Anatomization of Clustering of Iris Data Implemented using Genetic Algorithm

Sajid Saleem, Syed Nazeer Alam

***Abstract*—Genetic algorithms offer an approach to optimize the data. Genetics consist of genes which provide a blueprint to basic building block of life. Genes combined together to form a chromosome. Organisms share their genes, such as children share the genes of parents. This is called crossover. Permanent change can be done to the genes by means of mutation. Mutation is the process where a gene or a chromosome is permanently changed. Clustering algorithms attempt to revamp the positioning of like objects into homogeneous classes and objects. This paper describes the concept of genetic algorithm and clustering on Fishers IRIS to illustrate the concept of obtaining fittest gene from a given set of chromosomes. We employed K-means technique for clustering. We have generated chromosomes using IRIS data. In order to randomize and get unique chromosomes we applied crossover and mutation methodology repeatedly in cyclic manner to obtain optimized chromosomes. We tested each chromosome with a fitness function and seek optimized genes from the tested sample. Clustering applications are in various fields including marketing, biology, etc and genetic algorithms provide stochastic optimization techniques. Merging both technique help in efficient classification of biological samples, marketing samples, etc.**

***Keywords*— Clustering, Genetic Algorithm, Fitness, Chromosome**

1. INTRODUCTION

Clustering may be defined as “Given a set of data points, partition them into a set of groups which are as similar as possible”. One of the application areas of clustering include, but are not limited to, data mining problems, collaborative filtering, trend detection, data analysis and network analysis. There are many techniques available for clustering but the one adopted in our research is dependent on Distance-based algorithms called K-means because it is computationally faster and can produce tighter clusters than hierarchical clustering.

In K-means, the partitioning representatives communicate to the mean of each cluster. Distances are computed using the Euclidean distance formula[1], where the smaller distances components are grouped together to form a cluster, signifying that the smaller the distance between two data sets within a cluster the greater the identical compatibility. If the available data is only unlabeled, it is called as unsupervised clustering [2].

Constrained and unconstrained problem optimization solving imitating biological fruition based on natural selection is called Genetic Algorithm (GA). GA helps in achieving optimization techniques [3]. Biologically inspired operators like crossover as shown in Figure-1, and mutation as illustrated in Figure-2, are applied on chromosomes (encoded strings) to obtain a new generation of chromosomes. The processes of GA continue for certain duration of time until an exit condition is met. The common processes are selection, crossover and mutation [4].The phenomenon of clustering using genetic algorithm is explained as follows.

a. Clustering using Genetic Algorithm

The crossover and mutation capability of GA has been used in this article for the purpose of appropriately determining a fixed number cluster centers in the given data set, to finally suitably cluster the set of ‘n’ unlabeled data points. The sum of the Euclidean distances of the points is used as the clustering metric. The task is to obtain the lowest possible fitness value signifying the closeness / similarity of items within a cluster.

This research work carried out is important from the holistic point of view defining the common features of multiple species in biological domain, obtaining the latest trends with universal features in the global world of marketing.

II. RELATED WORK

To study related work is always source of inspiration for researcher to progress in the focused area. It is worth to mention that Whitley [5].In his work covers many genetic algorithms including validated as well as experimental genetic algorithms. The author’s basic objective is to provide a survey on genetic algorithms in such a way that a new comer to the field may get the gist of things, and at the same time an experience reader may easily absorb the concepts.

Bezdek et al.[6] focuses unsupervised clustering in their research and provide an approach for optimizing those using genetic algorithms. The authors use IRIS data

Manuscript Received July 19, 2016; accepted 18th November, 2016; date of current version December 2016

Sajid Saleem is with Electronics &Power Engineering Department, Pakistan Navy Engineering College
National University of Sciences & Technology (email: sajidsaleem@pnec.nust.edu.pk)
Syed Nazeer Alam is Electronics &Power Engineering Department, Pakistan Navy Engineering College
National University of Sciences & Technology (email: nazeer@pnec.nust.edu.pk)

set and apply03 different distance metrics for evaluation and optimization. The metrics include Euclidean, Mahalanobis and Diagonal techniques. The results are evaluated against hard c-means using the 03 mentioned metrics.

Jiang et al.[7] present an Integer Genetic Algorithm (IGA)for the analysis of cluster related problems. The authors propose 03 new genetic operators other than crossover and mutation namely; competition, self-reproduction and diversification. They also introduced a new clustering criterion and compared it with square-error criterion using real chemical data.

Hall et al.[7] in their work provide a genetically guided optimizing approach focusing hard means and fuzzy c-means in cluster analysis. They use 06 data sets including IRIS for their experimentation. They also show the time cost of genetic guided clustering.

Bezdek et al. [8], in their paper research the IRIS data set asto filtering out the replicas and identifying the original Fisher IRIS data. The replicas floating in the research community were causing errors and noncompliance in the experimentation.

Steinbeck et al. [8], present a research in comparing different clustering techniques like K-means and agglomerative hierarchical clustering. Within K-means they also explore a standard and bisecting algorithm. Their results show that bisecting K-means are better than standard k-means techniques and are either on par or better than hierarchical techniques.

Maulik and Bandyopadhyay[9], demonstrate the dominance of genetic algorithmbased clustering algorithm against k-means clustering. The 07 data sets used in the experiments include Vowel, IRIS and Crude Oil along with 04 artificially generated data sets.

Deb et al.[10], provide a non-dominated sorting genetic algorithm for multi-objective optimization (NSGA-II). Their results are compared with PAES, another similar type of genetic algorithm, and prove that their approach is better. This work is optimization of their earlier proposed Non-dominated Sorting Genetic Algorithm (NSGA) algorithm by Debb and Srinivas.

Garai and Chaudhuri [11], propose a genetic clustering algorithm that utilizes Hierarchical Clustering Merging Algorithm(HCMA) and Adjacent Cluster Checking Algorithm (ACCA).Their approach breaks the data into multiple fragments andthen checks them through algorithms. Their proposed techniqueis a type of split-merge based method which theycompare with the same type of algorithms (CURE, DBScanand Chameleon) in their experiments.

III. RESEARCH METHODOLOGY

Based on the facts discussed above we have opted the research methodology of K-means clustering for better results as mentioned in the introduction. The data set used for the experimentation is Iris flower data set also known as Fishers Iris data [9],. The dataset covers 03 species of IRIS namely setosa, virginica and versicolor. Fishers IRIS 50 samples of each group making it a total of 150 samples. The length and the width of sepals and petals in centimeters are calculated for each sample making 04 features.

We initiated by taking a random value between 0, 1, 2 to represent a row, hence taking 150 random values to represent each row. These 150 random values will form a chromosome. Once a chromosome is created it is grouped into clusters by using the concept of K-means clustering [10-13] and grouping similar values together, that is, groups of 0s, groups of 1s and groups of 2s, hence forming 03 clusters. After that fitness test is run through each cluster using Euclidean distance formula with the objective to obtain the least fitness value, signifying the best gene.
We are working in a two dimensional data set, we used 2D Euclidean distance formula mention in Eq-1 to measure the fitness of each gene within a chromosome.

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad (1)$$

We have repeated the mentioned procedure to obtain 10 chromosomes, each consisting of 03 clusters. Then we applied the process of crossover to all the chromosomes. The crossover process took first half values of chromosome 2 and merged them with the last half values of chromosome 1, and hence generated a new chromosome as depicted in Figure-1. The process continued on until the first half of chromosome 1 merges with the last half of chromosome 10. This results in 10 new chromosomes.

Algorithm: Crossover of Chromosomes

```
for $i$ = 1 to 9 do
    Take first half values of $x_{i+1}$ chromosome and merge them with last half of $x_i$ chromosome
end for
Take first half values of $x_{10}$ chromosome and merge them with last half of $x_1$ chromosome
```

The process of mutation is then applied on the samples to obtain further distinctive chromosomes. In mutation we randomly swapped 10 genes within a chromosome and then repeated the process, making it 02 times swap of 10 genes as illustrated in Figure-2.

This we carried out in order to break any relation of the new chromosome with the original one. The mutation process is applied on all 10 chromosomes that have already been passed through the crossover process.

Algorithm: Mutation of Chromosomes

```
while Repeat twice do
    Take 02 random genes within a chromosome
    Swap them
end while
```

The newly obtained 10 chromosomes again repeat the process of forming clusters and calculating fitness values. The chromosomes are again passed through crossover and mutation process and the cycle is repeated a total of 30 times. Finally we compare the fitness values of first 03 chromosomes obtained in each iteration and compare them to find the cluster with the best fitness value, which will in-turn give the best gene among the tested sample.

RESULTS AND DISCUSSION

After applying the techniques mentioned in the research methodology we obtain the results of the experimentation. The results of the 03 clusters of first 03 chromosomes from the sample of 10 chromosomes for all 30 iterations are comparable. Figure-3 illustrates that the gene obtained from cluster 01 of chromosome 03 obtains the best fitness value in round 12.Figure-4 show that the gene obtained from cluster 02 of chromosome 01 obtains the best fitness value in round 30. From Figure-5 we could conclude that cluster 03 has 03 best genes to be found at round 09 in chromosome 02, round 10 in chromosome 01 and round 24 in chromosome 03.The results obtained assist us to establish a better fitness function among the Fischer's IRIS data set. Other related work, as mentioned in section 2, explores the concepts of clustering and genetic algorithms with other datasets investigating aspects not limiting to fitness function.

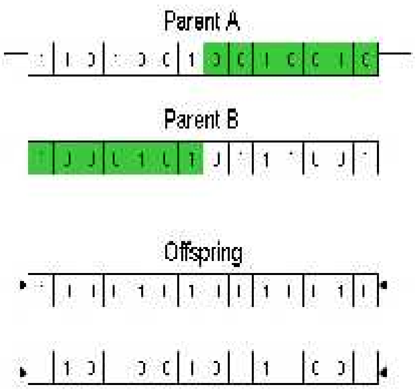

Figure.1: Sample Crossover Operation [4]

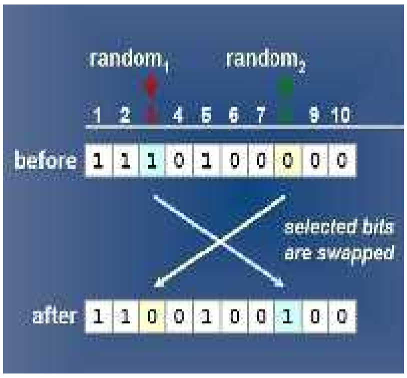

Figure.2: Sample Mutation Operation [5]

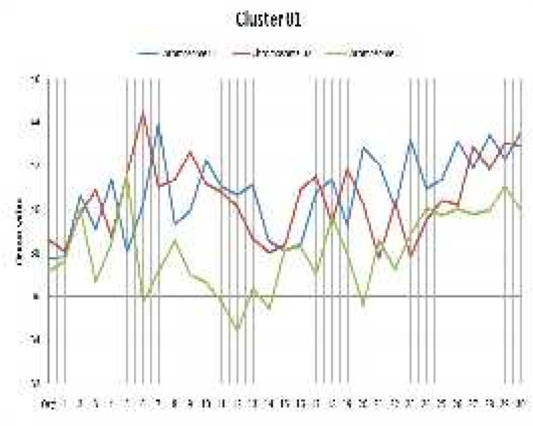

Figure.3: Comparison of cluster 01 of the first 03 chromosomes for all the 30 samples

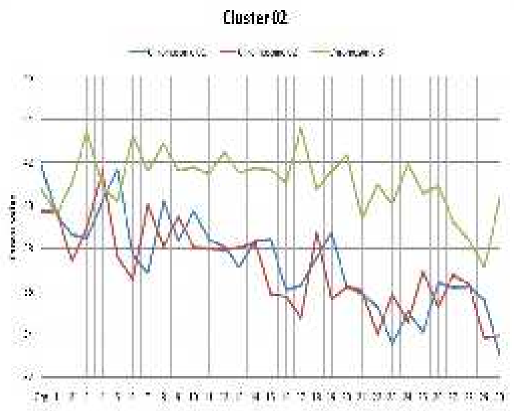

Figure.4: Comparison of cluster 02 of the first 03 chromosomes for all the 30 samples

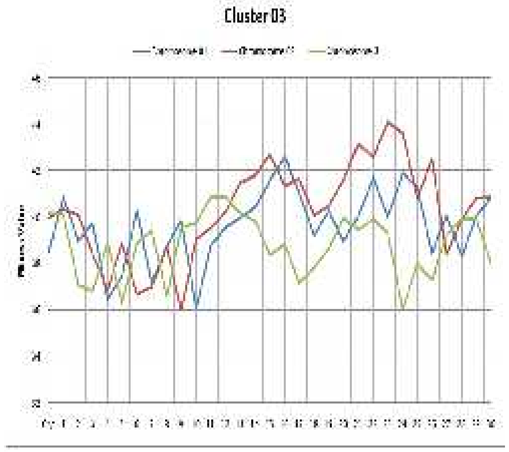

Figure.5: Comparison of cluster 02 of the first 03 chromosomes for all the 30 samples

CONCLUSION

We have evaluated the Fishers IRIS data merging them with the concept of genes and chromosomes in the domain of genetic algorithms. We applied the K-means clustering technique with a fitness function based on Euclidean distance to observe the results. The best genes among the chromosomes are the ones with the best fitness function value. The results have shown that we can further use the techniques to explore in various fields by simulating results pertaining to a particular problem. The future work includes enhancement of the algorithm to calculate and display the best chromosome among the sample of chromosomes. Also comparison among different data sets and among different fitness functions are some of the proposed futuristic research topics. We hope that the outcome of our effort will be appreciated and facilitated further in augmentation in the domain of interest.

ACKNOWLEDGEMENT

To carry out this research work we acknowledge for the provision of Fishers' IRIS data set from J. C. Bezdek et al., 1999 and related work carried out by the researchers.

REFERENCES

[1] Aggarwal, Charu C., and Chandan K. Reddy.2013,"Data Clustering: Algorithms and Applications". CRC Press.

[2] Bezdek, James C., Srinivas Boggavarapu, Lawrence O. Hall, and Amine Bensaid.,1994, "Genetic Algorithm Guided Clustering." In Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, IEEE, pp.34–39.

[3] Bezdek, James C., James M. Keller, Raghu Krishnapuram, Ludmila I. Kuncheva, and Nikhil R. Pal.1999, "Will the Real Iris Data Please Stand Up?" IEEE Transactions on Fuzzy Systems 7, no. 3: pp.368–69.

[14] Deb, Kalyanmoy, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan.2000, "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II." In Parallel Problem Solving from Nature PPSN VI, 849–58. Springer.

[5] GA Mutation Operator. Accessed June 19, 2016. http://www.edc.ncl.ac.uk/highlight/rhjanuary2007g04.php.

[6] Garai, Gautam, and B. B. Chaudhuri, 2004, "A Novel Genetic Algorithm for Automatic Clustering." Pattern Recognition Letters 25, no. 2, pp.173–87.

[7] Genetic Recombination, Mating and Gene Pairing. Accessed June 19, 2016. http://mnemstudio.org/genetic-algorithms-recombination.htm.
Goldberg, David E. 2006, "Genetic Algorithms". Pearson Education India.

[8] Hall, Lawrence O., Ibrahim Burak Özyurt, and James C. Bezdek. 1999,"Clustering with a Genetically Optimized Approach." Evolutionary Computation, IEEE Transactions on 3, no. 2 ,pp.103–12.

[9] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn, 1999, "Data Clustering: A Review." ACM Computing Surveys (CSUR) 31, no. 3, pp.264–323.

[10] Jiang, Jian-Hui, Ji-Hong Wang, Xia Chu, and Ru-Qin Yu.,1997, "Clustering Data Using a Modified Integer Genetic Algorithm (IGA)." Analytica Chimica Acta 354, no. 1–3 ,pp.263–74.

[11] Maulik, Ujjwal, and Sanghamitra Bandyopadhyay.,2000,"Genetic Algorithm-Based Clustering Technique." *Pattern Recognition* 33, no. 9 ,pp.1455–65.Mitchell, Melanie. *An Introduction to Genetic Algorithms*. MIT press, 1998.

[12] Steinbach, Michael, George Karypis, Vipin Kumar, and others, 2000,"A Comparison of Document Clustering Techniques." In *KDD Workshop on Text Mining*, vol.400. Boston, pp. 525–26.

[13] Whitley, Darrell, 1994, "A Genetic Algorithm Tutorial", Statistics and Computing 4, no. 2,pp.65–85.