Advantages and Limitations of Genetic Algorithms for Clustering Records

A. H. Beg

School of Computing and Mathematics
Charles Sturt University
Panorama Avenue, Bathurst 2795, Australia
E-mail: abeg@csu.edu.au

Md Zahidul Islam

School of Computing and Mathematics
Charles Sturt University
Panorama Avenue, Bathurst 2795, Australia
E-mail: zislam@csu.edu.au

Abstract—**Clustering is a fundamental and widely used method for grouping similar records in one cluster and dissimilar records in the different cluster. In cluster analysis, a major problem is to determine the appropriate number of cluster in advance. It is difficult for a user (data miner) to estimate the appropriate number of clusters in advance. Another limitation of a well-known clustering technique called K-means is that it gets stuck at local optima. In order to overcome these limitations Genetic Algorithm (GA) based clustering techniques have been proposed in the 1990s. Since then many researchers have developed several evolutionary algorithm based clustering techniques, including GA and applied in various fields. This paper presents an up-to-date review of some major GA-based clustering techniques for the last twenty (20) years. A total of 45 ranked (i.e. based on citation reports and JCR/CORE rank) GA-based clustering approaches are reviewed, which are uses for real-life applications such as real-life data sets, highway construction projects, a Gas Company, cellular networks and satellite image segmentations. Almost two third of the techniques do not require any user to define the number of clusters. Finally, a thorough discussion and emerging research directions are presented.**

Keywords—clustering; genetic algorithms; survey; evolutionary algorithms, literature review

I. INTRODUCTION

Clustering is a process of grouping records in such a way that the records in one cluster are very similar to each other than the records in different clusters [1, 2, 3]. In the past, many clustering methods have been proposed, and mostly these methods are broadly divided into two categories: hierarchical clustering [4-6] and partitioning clustering [8, 9, 10, 11, 19, 55]. Hierarchical clustering merges smaller cluster into a larger cluster or successively split the larger cluster into smaller clusters [1, 3].

A tree structure dendrogram is used to illustrate the hierarchical clustering. There are two types of hierarchical clustering namely agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering is the bottom-up approach where each individual record of the data set is considered as a cluster, and it iteratively merges two similar clusters. It uses some similarity measure for merging agglomerative clustering. The merging process is continued until getting a single cluster or the termination condition is satisfied [1, 3].

Divisive hierarchical clustering is a top-down approach where the records of the data set are considered as one large cluster. The large clusters are then split into a smaller cluster in such a way that the most similar records are placed in one cluster. The division process is continued until each record of the data set forms a separate cluster [1, 3]. The divisive hierarchical process is a kind of reverse process of the agglomerative hierarchical process. The main drawback of many hierarchical clustering is that they require high computational complexity. In general, the complexity of agglomerative hierarchical clustering is $O(n^3)$ and the complexity for divisive clustering is $O(n^2)$ [50], whereas the complexity of K-means (a partitioning clustering) is $O(n)$.

The goal of a partitioning clustering is to determine a partition $P = \{C_1, C_2, ..., C_k\}$ (i.e., $C_j \neq \emptyset, \forall k$; $C_i \cap C_j = \emptyset$; $\forall i \neq j$), where C_j stands for j-th cluster of partition P in such a way that the records which belong to same cluster are more similar to each other as possible whereas the records which belong to different cluster are as dissimilar as possible. K-means [2] is one of the most commonly used partitioning clustering algorithm. The main objective of K-means algorithm is to minimize the objective function that is described by the equation:

$$F(P) = \sum_{i=1}^{k} \sum_{x \in c_i} d(x, a_i) \qquad (1)$$

Where k stands for the number of clusters in the partition and a_i is the center of the cluster c_i, and $d(x, a_i)$ is the Euclidean distance between the record x and the cluster center a_i. The main purpose of this equation is to minimize the distance between each record and the center of the cluster. However, the most of the hierarchical and the partitioning clustering including K-means has a disadvantage: it is necessary to determine the number of cluster beforehand [8, 9, 10, 11, 60, 63]. In order to improve the performance of K-means various evolutionary algorithm based clustering approaches including genetic algorithms [8-14] have been proposed and applies in many different fields.

In this paper, we present a survey of Genetic Algorithms (GA) based clustering techniques for the last twenty (20) years, from 1995 to 2015 through a literature review. The major GA (i.e. based upon the number of citations and JCR/CORE rank) based clustering techniques during the period (1995-2015) are reviewed. This period is especially important because the GA-based clustering technique was first introduced in the 1990s and since then many GA-based clustering techniques have been proposed. A total of 45 GA-based clustering approaches are reviewed, which are uses in many real-life applications such as

real-life data sets, satellite image segmentation, DC-motor friction identification, cellular networks, map topologies, a gas company and highway construction projects and so on. Maximum of the techniques produces the clustering results without any user requirements.

The remaining part of the paper is organized as follows. In Section II, we discuss the genetic algorithms for clustering records. Section III discusses the limitations of various GA-based clustering approaches. Section IV discusses the advances of genetic algorithms based clustering techniques and in Section V, we provide the concluding remarks.

II. DISCUSSION ON GENETIC ALGORITHMS FOR CLUSTERING RECORDS

A. What is Genetic Algorithm?

Genetic algorithms (GA) are randomized search and optimization techniques that are motivated by the principals of natural selection and evolution processes [8, 9,10]. The GA was introduced by John Holland [15] in 1975. He was inspired by the concept of Darwinian's principle of survival of the fittest individuals and natural selection and developed the theory of genetic algorithm. Since then many researchers are using this algorithm to solve various optimization problems effectively [8-14].

B. Steps of Genetic Algorithms

In GA, the roles of initialization and recombination operators are very well defined. The initialization operator identifies the direction of search and recombination operator generates new regions for search [7]. GA first starts a generation with an *initial population*. The initial population is generated with a number of chromosomes. The chromosomes are made up with a number of *genes*. For clustering problem, a *chromosome* is considered to be a clustering solution, and a *gene* of a chromosome is considered to be the center of a cluster.

After initialization, in order to *selection,* an *objective/fitness* function is applied on each chromosome that identifies the goodness of a chromosome. Biologically inspired operators: *crossover* and *mutation* are then applied to the population in order to solve the clustering problem. At the end of each generation, GA applies an *elitist operation* where the newly generated populations are compared with the previous population. All these inter-related parameters and operators influence the performance of a GA [9, 11,12]. The processes of *selection, crossover, mutation* and *elitist operation* are continued for a fixed number of generations or until a termination condition is satisfied [7].

a) Step 1: Population initialization:

The first step of GA is to initialize population. To initialize population, size is one of the important things that need to be considered. The higher number of the population contributes towards a good clustering solution [16, 17]. Therefore, some GA-based clustering techniques [16, 28, 29, 30] uses big population for the clustering solution. Traditionally, the number of genes of a chromosome are generated randomly in the initial population. The records also selected randomly to form genes [9, 18, 19]. However, a careful selection of genes can create an initial population containing high-quality chromosomes. Having high-quality chromosomes in a population increase the possibility of getting better quality clustering solution at the end of genetic processing [8, 12, 20]. Therefore, some GA-based clustering techniques [8, 21] generate high-quality chromosomes in the initial population.

b) Step-2: Selection:

For the next genetic operator such as *crossover* and *mutation* GA selects chromosomes based on their *fitness/objective* function. There are various methods to calculate the fitness such as Davis-Bouldin (DB) Index [22], Sum of the Squared Error (SSE) [1, 18], Silhouette Coefficient [1] and COSEC [8].

c) Step-3: Crossover:

Crossover is an important step in GA where a pair of chromosomes swaps their segments/genes to each other and generates a pair of new offspring chromosomes. Typically, there are many selection criteria such as the roulette wheel [8, 19, 23, 24], rank-based wheel [18] and random selections [10] are used to select the chromosome pair for a crossover operation. Some GA-based clustering techniques [8, 21] use roulette wheel selection where the best chromosome (which is available in the current population) is chosen as one chromosome of the pair. The second chromosome of the pair is selected using the roulette wheel technique [19, 23, 24]. Blas et al. [18] use the rank-based wheel selection. In the rank-based wheel selection, the chromosomes are first sorted based on their quality and then the pair of chromosomes are chosen based on the position of a chromosome in the rank.

Once the pair of chromosomes is selected, GA then applies crossover operation on each pair of chromosomes. There are many approaches to performing crossover between a pair of chromosomes such as a single point [8, 19, 21, 25], multi-point [18], arithmetic [26], path-based [10] and heuristic [10]. In the single-point crossover, each chromosome of a pair is divided into two parts at a random point between two genes. The left part (having one or more genes) of one chromosome of a pair joins the right part of other chromosomes (having one or more genes) and form an offspring chromosome [8, 21]. In the multi-point crossover, each chromosome of a pair is divided into multiple parts and then swaps their part to each other and generates new offspring chromosome. In the path-based crossover [10] two parents, chromosomes create a path between them. Two points are then selected from the obtained path as offspring chromosomes. The heuristic crossover [10] uses the fitness values of two parent where the worst parent slightly moves towards the best parent.

d) Step- 4: Mutation:

Mutation randomly changes one or more genes of a selected chromosome with a probability equal to the mutation rate. There are many approaches for mutation such as such as division and absorption [9, 18, 19, 21], insertion [19], deletion [19], perturbation [19] and movement [19]. The division operation divides one cluster of a chromosome into two clusters. The absorption operation merges two clusters of a chromosome into one cluster. The perturb mutator randomly selects a cluster

center and changes the coordinates of the center [19]. The insert mutator randomly generates a center from the data set and inserts it into the chromosome [19]. The delete mutator deletes a randomly selected center of the chromosome [19]. The move mutator transfers one record from one cluster to another cluster and re-computes the cluster center of the chromosome [19].

e) Step-5: Elitist Operation:

The elitist operation [8, 9, 10, 21] preserves the best chromosome obtained so far at any stage (i.e. the iteration) of the GA and pass that on to the next generation in order to ensure that the best achievement so far does not get lost during genetic operations. If the fitness of the worst chromosome (P_w^i)) of the current (i^{th}) generation is less than the fitness of the best chromosome (P_b^{ALL}) found so far from previous all generations then the worst chromosome (P_w^i) is replaces with the best chromosome (P_b^{ALL}). Moreover, if the fitness of the best chromosome (P_b^i) of ith generation is greater than P_b^{ALL} then P_b^{ALL} is replaced by P_w^i.

III. LIMITATIONS OF GENETIC ALGORITHM-BASED CLUSTERING TECHNIQUES

In this section, we discuss some limitations of GA-based clustering techniques. There are some limitations of the existing GA-based clustering techniques. Many existing techniques [9, 34, 39, 47] randomly generate the number of genes in the chromosomes and the actual attribute values of the genes in the initial population/iteration. The quality of the genes is unlikely to be high due to the random selection process.

An existing called GenClust [8] generates high-quality chromosomes in the initial population and thereby obtains a good clustering result. However, the complexity of the initial population selection is very high *O(n²)*. Moreover, GenClust also requires a user to define different radii values for the clusters in the initial population selection. Therefore, one direction of research for GA-based clustering techniques can be the selection of high-quality chromosomes with low complexity.

Many GA-based clustering suffers from degeneracy. The degeneracy mainly occurs when multiple chromosomes represent the same solution [10]. Degeneracy can lead an inefficient solution as the same shapes of chromosomes are repeatedly occurred. In order to avoid the degeneracy, an existing technique called GAGR [10] introduces a gene-rearrangement approach. However, the gene-rearrangement approach used in GAGR requires the same size (i.e. the same number of genes) of pair chromosomes, which participates in crossover operation. Moreover, GenClust also uses a gene-rearrangement operation that can re-arrange the chromosome pair with different sizes. However, the gene-rearrangement used in GenClust can handle a dataset with low dimensions. Therefore, the techniques that can handle the gene-rearrangement for the data set with high dimension are desirable.

Moreover, the time complexities of some GA-based clustering techniques [8, 49, 53] are *O(nm² + n²m), O(n² + m²), O(n²)* respectively. Therefore, reducing the time complexity for GA-based clustering techniques is also highly desirable.

IV. APPLICATIONS OF GENETIC ALGORITHM-BASED CLUSTERING TECHNIQUES AND THEIR ADVANTAGES

In this section, we discuss the applications of GA-based clustering techniques and their advantages. Table 1 shows the list of major GA-based clustering techniques (i.e. based on the number of citations and JCR/CORE rank) from 1995 to 2015 and their applications. Casillas et al. [13] presented a GA-based clustering approach for document clustering. The proposed approach calculates an approximation of the optimum k value and solves the best grouping of the documents into k clusters. The algorithm experimented with sets of documents, which were the output of a query in a search engine.

Turgut et al. [14] used GA to enhance the performance of clustering technique in mobile ad hoc networks. They optimized their proposed weighted clustering algorithm (WCA). In their proposed technique, each cluster head handles the maximum possible number of mobile nodes in its cluster in order to facilitate the optimal operation of the medium access control (MAC) protocol. Consequently, it results in the minimum number of clusters and hence cluster heads.

A genetic c-means clustering algorithm (GCMA) [27] has been proposed, which is a hybrid technique combining the c-means clustering algorithm (CMA) with a genetic algorithm. GCMA is superior to CMA in the sense that it converges to a nearby global optimum rather than a local one. GCMA is applied to the problem of color image quantization. The experimental results show that substantial improvement of image quality is obtained by using the genetic approach.

Tzes et al. [31] used GA-based fuzzy clustering technique to provide a model of the friction in a dc-motor micro-maneuvering system. The proposed technique is fine-tuned through a genetic-based algorithm which adjusts the characteristics of the membership functions. The proposed model is used as a feedforward term for tracking purposes of the dc motor's angular velocity. The experimental result shows that the proposed scheme reduces the system's dead band, and compensates any hysteresis and/or disturbance-related effects.

Crowding clustering genetic algorithm (CCGA) [32] has been proposed, which combines crowding and clustering for multimodal function optimization, and analyze convergence properties of the algorithm. The proposed CCGA employs standard crowding strategy to form multiple niches and clustering operation to eliminate genetic drift. The numerical experiments on standard test functions indicate that the CCGA is superior to both standard crowding and deterministic crowding in quantity, quality, and precision of multi-optimum search. The CCGA is applied to the practical optimal design of varied-line-spacing holographic grating.

Xiao et al. [34] proposed a quantum-inspired genetic algorithm for k-means clustering called KMQGA. In the proposed KMQGA, a Q-bit based representation is employed for examination and exploitation in discrete 0 – 1 hyperspace using rotation operation of the quantum gate. The proposed technique uses typical genetic algorithm operators: selection, crossover, and mutation of Q-bits. KMQGA generates the optimal number of clusters and the optimal cluster centroids without knowing the exact number of clusters beforehand.

TABLE 1. LIST OF RANKED (BASED ON JCR/CORE) GENETIC ALGORITHM-BASED CLUSTERING TECHNIQUES OF THE LAST 20 YEARS AND THEIR APPLICATIONS

Authors	Year of publications	Number of citations	Rank (JCR/CORE)	Initial cluster number selection	Applications
R. Srikanth et al.	1995	88	Q2	Random	Real life datasets
D. Maio et al.	1995	24	Q2	Random	Map topologies
Murthy and Chowdhury	1996	284	Q2	Random	Real life datasets
P. Scheunders	1997	205	Q1	Random	Image segmentation
Tseng and Yang	1997	46	B	User defined	Real life data sets
A. Tzes et al.	1998	44	Q1	Random	DC-motor friction identification
Hanagandi and Nikolaou	1998	43	Q1	Random	Real life datasets
R. Cucchiara	1998	31	Q2	Random	Image segmentation
M. C. Cowgill et al.	1999	135	Q1	Random	Real life datasets
Lozano and Larranaga	1999	47	Q2	Random	Real life datasets
A. Demiriz et al.	1999	281	B	User defined	Real life datasets
Maulik and Bandyopadhyay	2000	1011	Q1	Random	Synthetic and real life datasets
Chiou and Lan	2001	108	Q1	User defined	Synthetic datasets
Tseng and Yang	2001	201	Q1	Random	Real life datasets
Dimopoulos and Mort	2001	75	Q2	Random	Cell-formation problems
C. H. Cheng et al.	2002	99	Q1	Random	Data partitioning
Bandyopadhyay and Maulik	2002	301	Q1	Random	Satellite image of a part of the city of Mumbai
D. Turgut et al.	2002	129	B	Random	Mobile ad hoc networks
Li and Chiao	2003	16	Q1	User defined	Image segmentation
Garai and Chaudhuri	2004	97	Q2	Random	Real life datasets
B. Abolhassani et al.	2004	15	Q1	Random	Cellular networks
M. K. Pakhira et al.	2005	160	Q1	User defined	Synthetic and real life datasets
J. C. Neto et al.	2006	67	Q2	Random	Leaf image segmentation
Laszlo and Mukherjee	2006	113	Q1	User defined	Real life datasets
S. Bandyopadhyay et al.	2007	182	Q1	Random	Image Segmentation
Laszlo and Mukherjee	2007	115	Q1	Random	Real life datasets
Hong and Kwong	2008	31	Q1	Random	Synthetic and real life datasets
L. Qing et al.	2008	62	Q1	User defined	Varied-line-spacing holographic gratings
W. Sheng et al.	2008	31	Q1	Random	Hand written signature data
D. Chang et al.	2009	88	Q1	User defined	Real life datasets
W. Song et al.	2009	75	Q1	Random	Real life datasets
S. Deng et al.	2010	31	Q1	User defined	Real life datasets
J. Xiao et al.	2010	42	Q1	Random	Real life datasets
Y. Liu et al.	2011	43	Q1	Automatic	Real life datasets
Yucenur and Demirel	2011	44	Q1	Random	Real life datasets
He and Tan	2012	41	Q2	Random	Real life datasets
L.E. A. Blas et al.	2012	52	Q1	Random	Synthetic and real life datasets
D. Chang et al.	2012	16	Q1	Random	Real life datasets
A. Aalaei et al.	2013	2	Q1	User defined	Mazandaran Gas Company in Iran
P. Festa	2013	6	Q3	User defined	Real life datasets
S. Mungle et al.	2013	15	Q1	Random	Highway construction project
S. Wikaisuksakul	2014	13	Q1	Random	Synthetic and real life datasets
Rahman and Islam	2014	12	Q1	Random	Real life datasets
P. Peng et al.	2014	4	Q1	Random	Real life datasets
T. P. Hong et al.	2015	2	Q1	Random	Real life datasets

A genetic algorithm-based efficient clustering technique called KGA clustering [35] has been proposed that utilizes the principles of K-means algorithm. The proposed technique exploits the searching capability of K-Means and avoids its major limitation of getting stuck at locally optimal values. The proposed technique is applied on classifying the pixels of a satellite image of a part of the city of Mumbai where the number of pixels is equal to 262,144 (512_512 image), which is a significantly large data to cluster.

Sheng et al. [36] developed a biometric authentication method which dynamically generates bio keys from the statistical features of biometric data. In their method, they employ a genetic algorithm to derive a nearly optimal quantization of the feature space via clustering. In the proposed biometric authentication method given it takes biometric samples as input and then a set of statistical features is first extracted from each sample. It then employed the fuzzy genetic clustering algorithm on each feature subset or single feature for model the intra and interuser variation. The proposed method is applied on handwritten signature data, which are considered to be representative of a near–worst case scenario within which the method can be evaluated. The experimental results show that the proposed approach to bio-key generation potentially yields good reliability and security in practice.

Aalaei et al. [37] present GA-based technique for creativity matrix clustering. Creativity is a promoting factor for any organization. Having employees with the creative manner increases the productivity of an organization. Their proposed method is equipped with a heuristic clustering technique and classifies the employees in order to creativity parameters and configures a creativity matrix. They applied clustering via mathematical programming. In order to solve the mathematical

model, they developed GA. The proposed method is applied in Mazandaran Gas Company in Iran.

Rahman and Islam [8] proposed a hybrid genetic algorithm with K-means called GenClust. The proposed technique is capable of automatically finding the right number of clusters and identifying the right genes through a novel initial population selection approach. The proposed technique also introduces a gene rearrangement operation that helps to produce high-quality cluster centers. The obtained cluster centers are then fed into K-means as initial seeds in order to produce an even higher quality clustering solution. The proposed technique is applied to some real-life datasets and shows encouraging results.

V. CONCLUSION

This paper presents an up-to-date review of some major GA-based clustering techniques for the last twenty (20) years (from 1995 to 2015). A total of 45 ranked (i.e. based on citation reports and JCR/CORE rank) genetic algorithm based clustering approaches are reviewed, which are uses for real-life applications such as real-life data sets, highway construction projects, a Gas Company, cellular networks and satellite image segmentations. Almost two third of the techniques do not require any user to define the number of clusters. The limitations of existing GA-based clustering techniques are described. We also present the applications of GA-based clustering techniques and their advances.

REFERENCES

[1] P.N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, first ed., Pearson Addison Wessley, 2005.

[2] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, pp. 129-13,1982.

[3] J. Han and M. Kamber, Data Mining Concepts and Techniques (2nd ed.). San Francisco: Morgan Kaufmann. 2006.

[4] J. Herrero, A. Valencia and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," Bioinformatics, vol. 17, pp. 126-136, 2001.

[5] P. Langfelder, B. Zhang and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," Bioinformatics, vol. 24, 719-720, 2008.

[6] C. F.Juang, C. M. Hsiao, and C. H. Hsu, "Hierarchical Cluster-Based Multispecies Particle-Swarm Optimization for Fuzzy-System Optimization," IEEE Transactions On Fuzzy Systems,vol. 18, 14-26, 2010.

[7] R.H. Sheikh, M.M. Raghuwanshi and A.N. Jaiswal, "Genetic algorithm based clustering: a survey," First International Conference on Emerging Trendsin Engineering and Technology, Nagpur, India, vol. 2, pp. 314-319, 2008.

[8] M. A. Rahman, & M. Z. Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-Means," Knowledge-Based Systems, vol. 71, pp. 345-365, 2014.

[9] Y. Liu, X. Wu and Y. Shen, "Automatic clustering using genetic algorithms," Applied Mathematics and Computation, vol. 218, pp. 267-1279, 2011.

[10] D. Chang, X. Zhang and C. Zheng, "A genetic algorithm with gene rearrangement for K-means clustering," Pattern Recognition, vol. 42, pp.1210-1222, 2009.

[11] M. Laszlo and S. Mukherjee, "A genetic algorithm that exchanges neighboring centers for k-means clustering," Pattern Recognition Letters,vol. 28, pp. 2359-2366, 2007.

[12] P. A. D. Gomez, D. F. Hougen, "Initial Population for Genetic Algorithms: A Metric Approach," In proceeding of: Proceedings of the 2007 International Conference on Genetic and Evolutionary Methods, GEM 2007, June 25-28, 2007, Las Vegas, Nevada, USA.

[13] A. Casillas, M. D. Lena and R. Martínez, "Document clustering into an unknown number of clusters using a genetic algorithm," Text, speech and dialogue, Springer-Verlag Berlin Heidelberg, LNAI vol. 2807, pp. 43-49, 2003.

[14] D. Turgut, S. K. Das, R. Elmasri and B. Turgut, "Optimizing Clustering Algorithm in Mobile Ad hoc Networks Using Genetic Algorithmic Approach", Global Telecommunications conference, 2002, IEEE GLOBECOM'02, vol.1, pp. 62-66.

[15] J.H. Holland, Adaptation in Natural and Artificial Systems, Univ. of Michigan Press, Ann Arbor, MI. 1975.

[16] J. Strasburga, C.G. Martelb andV. Alexandrov, "Parallel genetic algorithms for stock market trading rules," Procedia Computer Science, vol. 9, pp. 1306 -1313, 2012.

[17] H. Pourvaziri and B. Naderi, "A hybrid multi-population genetic algorithm for the dynamic facility layout problem," Applied Soft Computing, vol. 24, pp. 457- 469, 2014.

[18] L. E. A. Blas, S. S. Sanz, S. J. Fernández, L. C. Calvo, J. D. Ser and J. A. P. Figueras, "A new grouping genetic algorithm for clustering problems," Expert Systems with Applications. Vol. 39, pp. 9695- 9703, 2012.

[19] D. Chang, Y. Zhao, C. Zheng and X. Zhang, "A genetic clustering algorithm using a message-based similarity measure," Expert Systems with Applications, vol. 39, 2194-2202, 2012.

[20] D. E. Goldberg, V. Deb and K. J. Clark, "Genetic Algorithms, Noise, and the Sizing of Populations," Complex Systems, vol. 6, pp. 333-362,1992.

[21] A.H. Beg, & M.Z. Islam, "Clustering by Genetic Algorithm-High Quality Chromosome Selection for Initial Population," 10th IEEE conference on Industrial Electronics and Applications, ICIEA 2015, Auckland, New Zealand. 2015, pp. 129-134.

[22] D. L. Davies and D.W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell, vol. pp.1 224-227, 1997.

[23] A. Mukhhopadhyay and U. Maulik, "Towards improving fuzzy clustering using support vector machine: Application to gene expression data," Pattern Recognition, vol. 42, pp. 2744-2763, 2009.

[24] S. Bandyopadhyay, U. Maulik and M. K. Pakhira, Clustering using simulated annealing with probabilistic redistribution, International Journal of Pattern Recognition and Artificial Intelligence, vol. 15 269-285, 2001.

[25] P. Peng, O. Addam, M. Elzohbi, S. T. Özyer, A. Elhajj, S. Gao, Y. Liu, T. Özyer, M. Kaya, M. Ridley, J. Rokne and R. Alhajj, "Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data," Knowledge-Based Systems, vol. 56, pp. 108-122, 2014.

[26] X. Yana, Y. Zhua,W. Zoua, L and Wang, "A new approach for data clustering using hybrid artificial bee colony algorithm," Neurocomputing, vol. 97, pp. 241-250, 2012.

[27] P. Scheunders, "A Genetic C-Means Clustering Algorithm Applied To Color Image Quantization", Pattern Recognition, vol. 30, No. 6, pp. 859-866, 1997.

[28] Y. Y. Liu, and S. Wang, "A scalable parallel genetic algorithm for the Generalized Assignment Problem," Parallel Computing, vol. 46, p. 98-119, 2015.

[29] M. Moore, "An accurate parallel genetic algorithm to schedule tasks on a cluster," Parallel Computing, vol. 30, pp. 567-583, 2004

[30] J. Kumar, R. T. Richard Mills, F. M. Hoffman and W. W. Hargrove, "Parallel k-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets," Procedia Computer Science, vol. 4, pp. 1602-1611, 2011.

[31] A.Tzes, P. Y. Peng, J. Guthy, "Genetic-Based Fuzzy Clustering for DC-Motor Friction Identification and Compensation," IEEE Transactions on Control Systems Technology, vol. 6, no. 4, 1998.

[32] L. Qing, W. Gang, Y. Zaiyue, W. Qiuping, "Crowding clustering genetic algorithm for multimodal function optimization," Applied Soft Computing, vol.8, pp. 88-95, 2008.

[33] J. C. Neto, G. E. Meyer, D. D. Jones, "Individual leaf extractions from young canopy images using Gustafson–Kessel clustering and a genetic algorithm," Computers and Electronics in Agriculture, vol. 51, pp. 66-85, 2006.

[34] J. Xiao, Y. P. Yan, J. Zhang, Y. Tang, "A quantum-inspired genetic algorithm for k-means clustering," Expert Systems with Applications, vol. 37, pp. 4966 - 4973, 2010.

[35] S. Bandyopadhyay and U. Maulik, "An evolutionary technique based on K-Means algorithm for optimal clustering in R^N", Information Sciences, vol. 146, pp. 221-237, 2002.

[36] W. Sheng, G. Howells, M. Fairhurst, and F. Deravi, "Template-Free Biometric-Key Generation by Means of Fuzzy Genetic Clustering," IEEE Transactions on Information Forensics and Security, vol. 3, no. 2, 2008.

[37] A. Aalaei, H. Fazlollahtabar, I. Mahdavi, N. M. Amirid and M. H. Yahyanejade, "A genetic algorithm for a creativity matrix cubic space clustering: A case study in Mazandaran Gas Company," Applied Soft Computing, vol. 13, pp.1661-1673, 2013.

[38] R, Srikanth, R, George, N, Warsi, D, Prabhu, FE, Petry and BP, Buckles, "A variable-length genetic algorithm for clustering and classification," Pattern Recognition Letters, vol. 16, no. 8, pp. 789-800,1995.

[39] D, Maio, D, Maltoni and S, Rizzi, "Topological clustering of maps using a genetic algorithm," Pattern Recognition Letters, vol. 16, no.1, pp. 89-96, 1995.

[40] CA, Murthy and N, Chowdhury, "In search of optimal clusters using genetic algorithms," Pattern Recognition Letters, vol. 17, no. 8, pp. 825-32, 1996.

[41] L. Y. Tseng and S. B. Yang, "Genetic algorithms for clustering, feature selection and classification," in Proc. IEEE Int. Conf. Neural Networks, New York, vol. 3, pp.1612-1615,1997.

[42] V. Hanagandi, M. A. Nikolaou, "A hybrid approach to global optimization using a clustering algorithm in a genetic search framework," Computers Chem. Engng, vol. 22, no. 12, pp.1913-1925, 1998.

[43] R. Cucchiara," Genetic algorithms for clustering in machine vision," Machine Vision Application. , vol. 11, no. 1, pp. 1-6, 1998.

[44] M. C. Cowgill, R. J. Harvey, and L. T. Watson, "A genetic algorithm approach to cluster analysis," Comput. Math. Appl., vol. 37, pp. 99-108, 1999.

[45] J. A. Lozano and P. Larranaga, "Applying genetic algorithms to search for the best hierarchical clustering of a dataset," Pattern Recognit. Lett., vol. 20, no. 9, pp. 911–918, 1999.

[46] A. Demiriz, K. Bennet, and M. Embrechts. Semi-supervised clustering using genetic algorithms. In Proceedings of Artificial Neural Networks In Engineering, 1999.

[47] U. Maulik and S. Bandyopadhyay, "Genetic Algorithm Based Clustering Technique," Pattern Recognition, vol. 33, pp. 1455-1465, 2000.

[48] Y. C. Chiou and L. W. Lan, "Theory and methodology genetic clustering algorithms," Eur. J. Oper. Res., vol. 135, no. 2, pp. 413–427, 2001.

[49] L. Tseng and S. Yang,"A genetic approach to the automatic clustering problem,"Pattern Recognit., vol. 34, pp. 415-424, 2001.

[50] C, Dimipoulos and N. A, Mort, "Hierarchical Clustering Methodology Based on Genetic Programming for the solution of simple Cell-Formation Problems," International Journal of Production Research, vol. 39, pp. 1-19, 2001.

[51] C. H. Cheng, W. K. Lee, and K. F. Wong, "A genetic algorithm-based clustering approach for database partitioning," IEEE Trans. Syst. Man Cybern. B, vol. 32, pp. 215–230, Aug. 2002.

[52] C.T. Li and R. Chiao, "Multiresolution genetic clustering algorithm for texture segmentation," Image and Vision Computing, vol. 21 no. 11, pp. 955–966, Oct. 2003.

[53] Garai, G. and Chaudhuri, B. B., "A novel genetic algorithm for automatic clustering," Pattern Recognition Letters, Vol.25, Issue2, pp. 173-187, 2004.

[54] B. Abolhassani, J. E. Salt, and D. E. Dodds, "A two-phase genetic k-means algorithm for placement of radio ports in cellular networks," IEEE Trans. Syst., Man, Cybern. B, Cybern.,vol. 34, pp. 533–538, 2004.

[55] W. M. B. W. Mohd, A.H. Beg, T. Herawan and K. F. Rabbi, "An Improved Parameterless Data Clustering Technique based on Maximum Distance of Data and Lioyd k-means Algorithm," Procedia Technology, Vol. 1, pp. 367-371, 2012.

[56] M. Pakhira, S. Bandyopadhyay, and U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classi-fication," Fuzzy Sets Syst., vol. 155, pp. 191–214, 2005.

[57] M. Laszlo and S. Mukherjee, "A Genetic Algorithm Using Hyper-Quadtrees for Low-Dimensional K-Means Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 533-543, Apr. 2006.

[58] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 5, pp. 1506-1511, 2007.

[59] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, "Unsupervised feature selection using clustering ensembles and population-based incremental learning algorithm," Pattern Recognit., vol. 41, no. 9, pp. 2747-2756, 2008.

[60] W. M. B. W. Mohd, A.H. Beg, T. Herawan and K. F. Rabbi, "MaxD K-Means: A Clustering Algorithm for Auto-generation of Centroids and Distance of Data Points in Clusters," Computational Intelligence and Intelligent Systems (ISICA), Communications in Computer and Information Science (CCIS 2012). pp. 192-199, 2012.

[61] W. Song, C. H. Li, and S. C. Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures," Expert Systems with Applications, vol. 36, pp. 9095-9104, 2009.

[62] S. Deng, Z. He, and X. Xu, "G-ANMI: A mutual information based genetic clustering algorithm for categorical data," Knowledge-Based Systems, vol. 23, no. 2, Mar. 2010, pp. 144-149, 2009.

[63] G. N. Yucenur and N. C. Demirel, "A new geometric shape-based genetic clustering algorithm for the multi-depot vehicle routing problem," Expert Systems with Applications, vol. 38, pp. 11859-11865, 2011.

[64] W. M. B. W. Mohd, A.H. Beg, T. Herawan, A. Noraziah and K. F. Rabbi, "Improved Parameterless K-Means: Auto-Generation Centroids and Distance Data Point Clusters," International Journal of Information Retrieval Research (IJIRR). Vol. 1, no. 3, pp. 1-14, 2011.

[65] H. He and Y. Tan, "A two-stage genetic algorithm for automatic clustering," Neurocomputing, vol. 81, pp. 49–59, Apr. 2012.

[66] P. Festa, "A biased random-key genetic algorithm for data clustering," Mathematical biosciences, vol. 245, pp. 76-85, 2013.

[67] S. Mungle, L. Benyoucef, Y.-J. Son, and M. Tiwari, "A fuzzy clustering-based genetic algorithm approach for time–cost–quality trade-off problems: A case study of highway construction project," Engineering Applications of Artificial Intelligence, 2013.

[68] S, Wikaisuksakul, "A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering," Applied Soft Computing, vol.24, pp. 679-91, 2014.

[69] T-P, Hong, C-H, Chen, F-S, Lin, "Using group genetic algorithm to improve performance of attribute clustering," Applied Soft Computing, vol. 29, pp. 371-8, 2015.