PERFORMANCE EVALUATION OF NOVEL SELECTION PROCESSES THROUGH HYBRIDIZATION OF K-MEANS CLUSTERING AND GENETIC ALGORITHM

HAQ, E. U.[1] – HUSSAIN, A.[4] – AHMAD, I.[2,3,1]

[1]*Department of Mathematics and Statistics, Faculty of Basic and Applied Sciences, International Islamic University, 44000 Islamabad, Pakistan*

[2]*Department of Mathematics, King Khalid University, Abha 62529, Kingdom of Saudi Arabia*

[3]*Statistical Research and Studies Support Unit, King Khalid University, Abha 62529, Kingdom of Saudi Arabia*

[4]*Department of Statistics, Quaid-e-Azam University, 44000 Islamabad, Pakistan*

**Corresponding author*
e-mail: abid0100@gmail.com; phone: + 92-33-3431-0207

(Received 30th May 2019; accepted 25th Oct 2019)

Abstract. K-means clustering combined with genetic algorithm (GA) techniques are used to improve the accuracy of estimation process and to minimize computational effort for solving nonlinear optimization problems. The main purpose of K-means clustering is to exhibit faster convergence which turns into quick evolution. This paper focuses on newly proposed cluster based GA selection techniques for solving unconstrained optimization problems. The K-means cluster based genetic algorithm (GKA) selection techniques comprise of four major stages: clustering, membership probability indexing, fitness evaluation and selection. The hybridization of genetic algorithm and clustering will effectively cater the problem of population diversity and selection pressure. There are two types of GKA selection techniques that are examined, the first selection technique (GKA_F) includes two proposed selection operators which are linked with a fixed number of clusters while the second technique (GKA_{opt}) is based on the optimum number of clusters. The main focus of these new selection techniques is to preserve population diversity as well as to avoid local optima. The performance of each technique is evaluated through eleven well known benchmark functions. On the whole, the novel cluster based selection techniques are demonstrated to be extremely efficient and effective for achieving optimum solutions which are verified by simulated results.
Keywords: *cluster evaluation, selection operators, benchmarks, selection pressure, population diversity, comparisons*

Introduction

There has been a significant growth in the fields of artificial intelligence, computational analysis, data mining and optimization in recent years. Classical techniques are unable to solve complex problems efficiently in the fields of computational engineering, transportation, energy, and management (Zhang et al., 2014). Hence the edification of optimization algorithms can be classified into stochastic and deterministic approaches (Fister et al., 2013). Mostly, deterministic algorithms are gradient based algorithms that employ the function values with their derivatives. These algorithms are very much useful for smoothing unimodal problems, but in terms of some discontinuous functions, non-gradient algorithms will be preferred (Yang, 2014). Hooke–Jeeves pattern and Nelder–Mead downhill simplex (Rajan and Malakar, 2015) search techniques are some of the examples of non-gradient based algorithms. In regard of stochastic approach, heuristics and meta-heuristics are two types of stochastic

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

algorithms. The major focus of stochastic techniques is to obtain feasible solutions at optimum scenario. There is no surety for finding absolute optimum solutions; however, it is presumed that mostly stochastic algorithms will achieve nearly optimal solutions.

Genetic algorithms (GAs) are stochastic based-heuristic search techniques that originated from biological evolution theory and applied in solving practical problems in the field of human developments. In the process of GAs, sometimes the optimal solution may not be feasible due to internal deficiencies such as less computational efficiency and premature convergence (Aibinu et al., 2016). GA does not have much mathematical requirements. Due to their developmental nature, GA will hunt down arrangements without looking to the particular internal working of the issue. It can deal with any sort of target capacities and imperatives (i.e. direct or nonlinear) characterized on discrete, non-stop or blended inquiry spaces. Hence the main focus of genetic algorithm (GA) is to find the best techniques through suitable adjustment between exploration (population diversity) and exploitation (selection pressure/premature convergences) (Haq et al., 2019a).

On the other hand, clustering is an algorithmic technique to organize the numeric information into meaningful groups. It can also be described as an unsupervised arrangement whereby the data values are clustered using specific information that is available in the dataset and also have prior knowledge about the number of clusters 'K'. This method is often used to discover the patterns of a given dataset (Li et al., 2015). The dataset contains information on variables and usually one attempt to reorganize the useful variables that have the same characteristics into the same group or cluster. However, the main challenge with the clustering is that different clustering algorithms may provide different clusters for the same dataset (Rehman and Islam, 2011). A good clustering algorithm is the algorithm that can reflect the natural clusters in a dataset and at the same time, attain the lowest validity index value (Islam et al., 2018). The clustering validity indices usually measure the compactness and the differentiability of the clusters. The detailed summary regarding strength and weakness of GA (Sivanandam and Deepa, 2008; Aibinu et al., 2016) and K means clustering (Islam et al., 2018) is presented in *Table 1*.

Genetic algorithm is one of the well-known optimization algorithms used to overcome K-means weakness. The major focus of GA based algorithm was to generate high quality clusters in minimum time. Some algorithms have also been designed in a multi-objective optimization form to understand and implement problems that are multifarious. In present study, we will focus on the cluster based selection techniques in genetic algorithm (GA), where clustering is used to organize the population of chromosomes/individuals for the process of reproduction and recombination. Hence, these newly proposed K-means cluster based genetic algorithm (GKA) selection techniques are effectively handle unconstraint optimization problems. The GKA method comprises of four major stages: clustering, membership probability indexing, fitness evaluation and selection. Hence the hybridization of genetic algorithm and clustering will effectively cater the problem of population diversity and selection pressure. A membership selection probability to each individual is followed by clustering. Fitness scaling modified the membership results in regard of selection function. Here two versions of (GKA) selection techniques are examined, the first selection technique (GKA_F) has two proposed selection operators which are linked with fixed number of clusters and the second technique (GKA_{opt}) is based on the optimum number of clusters. The performance of each technique is evaluated through eleven well-known benchmark

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

functions. The simulated results reveal that the proposed cluster based selection methods outperform as compared to others for achieving optimum solutions.

Table 1. *Detailed summary regarding strength and weakness of algorithms*

Type	Strength	Weakness
Genetic algorithm	• Conceptually easy to understand and execute • Efficiently perform for large scale complex optimization problems • Handle complex and noisy functions easily • Powerfully handle difficulties in evaluation process of the objective function • Require no prior knowledge or gradient information about the problem • Avoid to become stuck at local optima	• Difficulty in identifying fitness function and representation of optimization problems • Occurrence of premature convergence • Difficulty in selection of different parameters like population size, crossover and mutation rate etc. • Population diversity • Configuration is not so simple and straightforward
K-means clustering	• Relatively easy to implement • Computationally faster than other clustering methods with large population size • Clustering process is surely convergent • Easy adaptation for new examples • Easy generalization of clusters to different shapes and sizes	• Number of clusters must be determined before the iterative process begins • Clustering result is extremely sensitive to the initial seed-points • Noise, or outliers and empty clusters decline the superiority of the K-means clustering result • Neglects to recognize non-straight detachable groups in the input space

The remainder portion of this research is presented as: in “Literature review”, we comprise of some relevant study for K-means clustering co-integrated with optimization algorithms. Defining problem along with working strategy of genetic algorithm is described in “Defining problem through hybridization of K-means clustering and genetic algorithm”. The proposed selection strategy is comprehensively discussed in “Proposed K-means cluster based GA selection operators”. A detailed description about the benchmark functions is presented after the proposed work, while simulated results and performance evaluation of proposed methods are demonstrated in “Statistical results and discussion“. Finally, “Conclusions and future work” is provided at the end of the study.

Literature review

There are several algorithmic techniques like simulated annealing (Hatamlou et al., 2012) in the literature which are helpful to solve cluster problems. GA is one of the most promising algorithms that have consistently performed well in solving clustering problems. The capability of GA has proven to obtain efficient and effective results and provide appropriate clustering. In the past several years, GA has been extensively used as an optimization method in various domains such as image processing (Loai et al.,

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

2008; Younus et al., 2015; Huang and Ma, 2019; Belahbib and Souami, 2011), and clustering (Lin et al., 2005; Maulik et al., 2011; Murty et al., 2008) to name a few examples. There are several literature reviews that focus on the application of GAs to cluster integer data. All of these methods showed good performance and better results when compared to other clustering methods. However, some of these methods have drawbacks and need to be improved to develop a better clustering algorithm.

Genetic K-means Algorithm, GKA is one of the examples of GAs that was proposed to improve the performance of K-means. The main objective of GKA is to find the global optimum of the given dataset and partition the data into a specified number of clusters. In GKA, instead of using a common crossover operator K-means are used (Zeebaree et al., 2017) as search operators. The problem of minimizing the total within cluster variation (TWCV) was also handled successfully by GKA.

Lu et al. (2004) proposed a Fast Genetic K-means Algorithm, FGKA, which was inspired by GKA, by incorporating several improvements over GKA. Both FGKA and GKA achieved the objective of their studies which converged to the global optima, and the study found that FGKA runs faster than GKA. Maulik et al. (2011) proposed a GAs based clustering where chromosomes were represented by the strings of the real numbers and encoded a fixed number of cluster centers in RN. This algorithm was then extended by Maulik et al. (2011) and named as Genetic Clustering for Unknown K (GCUK). To check the performance of the algorithm, Maulik et al. (2011) compared the minimum value of the objective function in the K-means algorithm with the same K, and showed that the GCUK outperform the K-means. In GCUK, Maulik et al. (2011) used the Davies Bouldin (DB) index to measure the validity of the clusters. These two algorithms used the Euclidean distance to calculate the distance from a point to a cluster center. GCUK became the most effective GAs clustering method but due to the real number representation, it took a longer time to converge (Lin et al., 2005).

In a paper by Lin et al. (2005), the cluster centers were selected directly from the data set and they constructed the look-up table to save the distances between all pairs of the data points. This process allowed the algorithm to speed up the evaluation of the fitness value. In GCUK, GAs clustering by Maulik et al. (2011), the string representation was used to encode the variable number of cluster centers, while Lin et al. (2005) used the binary representation. A cluster based genetic algorithm with polygamy and dynamic population control procedure have been suggested by Aibinu et al. (2016) with an application of route optimization problem. Islam et al. (2018) presented an effective genetic algorithm that combines the capacity of genetic operators to conglomerate different solutions of the search space with the exploitation of the hill-climber cycles of K-means.

Defining problem through hybridization of K-means clustering and genetic algorithm

It is very important for any clustering algorithmic to find approximate or global optima for complex nonlinear optimization problems (Maulik et al., 2011). The K-means clustering algorithm is quite likely to converge to a suboptimal position. The key benefit of stochastic optimization approach over deterministic techniques is that they are unable to converge to local optima. Hence, stochastic techniques are able to solve clustering related problems; such as genetic algorithms, ant colony optimization, simulated annealing and other evolutionary techniques.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Genetic Algorithms (GAs) play with the idea of the evolutionary process where the chromosomes will have to compete with each other to have a place in the next generation. Strong chromosomes have more chance to survive and usually the weak chromosomes have limited chance. GAs are working with a search space that contains all feasible solution. It means that each of the points in the search space represents one feasible solution that will be marked according to its fitness through objective functions. The core processes of GAs are selection, crossover and mutation. All the processes of GAs make this algorithm more unique as compared to other conventional algorithms for the optimization. The selection process aims to select the good chromosomes which will be sent to the mating pool to combine with the other chromosomes where the features of parents are combined to form new offspring through crossover process (Haq et al., 2019b). Crossover process creates offspring with the help of those parents which are selected through the selection operator. Meanwhile the mutation process aims to encourage diversity in the new population with a very small probability.

Normally GA is a deliberately stochastic process, i.e. an ordered chain of well-defined states, whereby suitable solutions to given problems are determined by using a population of selected individuals. These individuals are examined by using some fitness criterion against the specific problem and using some predefined convergence or stopping procedure for the process to be terminated (Devooght, 2010). In Markov Chain, development of problem is based on modelling and analysis of GA (Eiben and Smit, 2011). Therefore, one of the most significant feature of GA is stochastic aspect which demonstrate the selection and formation of new chromosomes.

The conventional GA, may be characterized by the following schematic scheme:

1. Selection of all individuals as parents followed by a randomization approach (Zeebaree et al., 2018).
2. Genetic operators are used for a generation of offspring.
3. The new population of chromosomes are selected from the mixture of the old and newly generated offspring without changing the size of the string.
4. The main features including the fitness value of new chromosomes are examined against the termination criterion, the algorithm either stops or continue to the next step.

Genetic operators have the ability to maintain the genetic diversity throughout the generations. Variability in genetics or genetic diversity is necessary for the evolutionary process. The core intension of the genetic process is the creation of the fittest population which depends on the valuable cooperation between the genetic operators. The initiation of the idea regarding clustering in genetic algorithm is originated in the context to enhance the quality of solutions by avoiding excessive exploitation and restricting local optima instead of global optimum solution (local maximum or minimum solutions). The methodology of clustering is to enhance the selection probability for the convergence to the global optima by adequately covering the solution space, yet ensuring appropriate selection pressure to attain even better solutions from current population.

Moreover, there is not a rule of thumb for evaluating the performance of GA by choosing an appropriate optimization function. Therefore, the performance of the algorithm is based on the nature of the problem regarding variation rate in objective function and the number of local optima etc. (Hussain et al., 2017). A multimodal function has at least two local optima. The efficient search procedure must be proficient in eliminating the region around local optimum in context of the search for global

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

optima. The scenario becomes more complex in a situation of random distribution of local optima in search space. By hybridizing the strength of genetic algorithm and clustering of the fitness values, a detailed description of the proposed KGA methodology is revealed in the following section.

Proposed K-means cluster based GA selection operators

In this section, we briefly describe the process of proposed selection operators of genetic algorithm K-means clustering (GKA) for obtaining optimum solution of unconstraint optimization problems. Here, we use a standard genetic algorithm with objective function $f(\vec{x})$ to evaluate the performance of proposed GKA algorithm. These proposed techniques which cater the short comings associated the conventional section (selection) methods by minimizing the distance between centers and individuals by enhancing the search space.

GKA$_F$

The proposed methodology about selection operators of GKA is unique in such a way that the individuals of the population are divided into homogenous groups/clusters. These clusters are internally homogenous and externally heterogeneous as possible. These newly proposed selection techniques will resolve two important issues i.e. exploitation and exploration. Exploration means identifying potential areas of search space and discover new knowledge (Yang et al., 2009). It is also a process of attaining new information by visiting new states. Exploitation generates information and transmission of adaptation, which means optimizing within a promising region. Pure stochastic search is suitable at exploration while hill climbing is best at exploitation (Das et al., 2008). Recombination of the individuals within same cluster reduces population diversity and thus compromising scenario between the exploration and exploitation is mainly determined through clustering within individuals of the population.

The basic aim of the clustering is to explore a specific sequence among the data points that are exploratory in nature (Jain, 2010). The main focus is to organize datasets by using clustering technique which requires some divergence among the nature of the datasets and according to the purpose of analysis. Several types of clustering algorithmic techniques have been proposed (Jain, 2010; Belhaouari et al., 2014; Xu and Wunsch, 2005; Islam et al., 2018) i.e. taxonomy of clustering, discussions on primary short comings and major issues. One of the simplest and most popular clustering algorithmic technique is the K-means algorithm (KMA), and was originated by Steinhaus (1956). Although it was quite long ago, this technique is still the most widely used algorithm for clustering.

Our interest in this study was to cluster the fitness values to observe the same pattern for the selection of individuals. Illustratively, a set of '*n*' variables $X = \{x_1, x_2, x_3 \dots \dots x_n\}$ to be clustered with each of these $X_i \in R^p$is an attribute vector used to describe the variables. These variables will be clustered into a set of clusters, $C = \{C_1, C_2, C_3 \dots \dots C_k\}$ where K is the number of clusters.

The clusters are mutually exclusive $C, C_i \cap C_j = \emptyset \; for \; i \neq j$. The numbers of K may be priori known or not. Let $\bar{x}_k$ be the mean of the cluster C_k. The main objective of clustering is to find the minimum distance between x_i to the closest center $\bar{x}_k$ as follows in *Equation 1*.

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \bar{x}_k\|^2 \qquad \text{(Eq.1)}$$

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

The most familiar clustering technique is K-means (Islam et al., 2018) clustering which is an iterative procedure which is flexible to implement (Jain, 2010).

The procedure of K-means clustering along with inclusion of selection probabilities through GA can be described below:

1. Initially, the cluster centers $\{\bar{x}_1, \bar{x}_2, \bar{x}_3 \dots \dots \dots, \bar{x}_k\}$ are selected for each respective clusters $\{C_1, C_2, C_3 \dots \dots C_k\}$.
2. The new cluster membership is calculated by assigning each data value to the closest centers.
3. After allocation of data value to new cluster, cluster centers are re-computed.
4. If all cluster center stay in its position the algorithmic process is terminated. Otherwise, the procedure will be repeated from Step 2.
5. The selection probabilities are computed from proposed selection operators.
6. The parents are selected through recombination process.

To improve the performance of the GA process, a newly proposed cluster based selection operator through membership probability index that is assigned to each individual after clustering phase. Basically fitness evaluation will be transformed into membership probabilistic scores in range those are appropriate for selection. The membership probability score of an individual is a measurement of its affiliation with respect to both designated and external clusters. Hence proposed cluster based selection techniques create a balance between selection pressure and population diversity. In other words, these techniques are helpful for suitable adjustment between selection pressure and population diversity (Hussain and Muhammad, 2019). It has been also perceived that the mating pool may comprise of all higher proportionate individuals on absolute uniform scaling. These proposed selection techniques will be helpful in minimizing the selection pressure and improving the search space through clustering approach. The initiation of K-means clustering concept with new selection probability indices will definitely introduce greater diversity in the population thus offering better solution with sustainable convergence speed. In fact, new selection techniques create a balance between selection parameters. The mathematical ecology of the two newly proposed selection probability indices are shown in *Equations 2–7:*

Cluster based selection operator-1

$$P_{(i,j,l),1} = \frac{1}{N}\left[\frac{h_j}{N(h_j-1)}\left(\sum_{l=1}^{N} Z_l - NZ_l\right) + 1\right] \quad \text{(Eq.2)}$$

where

$$Z_l = \frac{f(x_i)}{h_j\left(e^{\sum_{i=1}^{h_j} \frac{\ln(f(x_i)}{h_j}}\right)} \quad \text{(Eq.3)}$$

and

$$N = \sum_{l=1}^{K} h_j \quad \text{(Eq.4)}$$

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Cluster based selection operator-2

$$P_{(i,j,l),2} = \left[\frac{W_j}{(h_j - 1)}\left(\sum_{l=1}^{N} Z_l - NZ_l\right) + \frac{1}{N}\right] \quad \text{(Eq.5)}$$

where

$$Z_l = \frac{f(x_i)\left(\sum_{i=1}^{h_j} \frac{1}{f(x_i)}\right)}{h_j^2} \quad \text{(Eq.6)}$$

and

$$W_j = \frac{h_j}{N} \quad \text{(Eq.7)}$$

where N is the size of the population of individuals, K is the number of cluster along with cluster size h_j. W_j is the proportion of the cluster from individuals' population.

Here some theoretical findings associated with cluster based selection probability are given below.

1. The cumulative probabilities of j^{th} clusters with size h_j is equal to W_j, hence clusters with more individuals will be obtained larger probability sum of each cluster. Moreover, an individual with higher fitness value within the cluster allocated lower selection probability to control selection pressure and population diversity increased.
2. By reducing the recombination probability, the cluster will avoid premature convergence and lower selection probability will be awarded to each individuals in larger clusters.
3. The cumulative selection probability is equal to one i.e. $\sum_{i=1}^{N} P_{(i,j,l),1} = \sum_{i=1}^{N} P_{(i,j,l),2} = 1$.

In the process of fitness evaluation, fitness scaling transforms the membership scores in a range appropriate for the selection function which selects the parents for the next generation. The selection function assigns a higher selection probability to individuals with higher scaled values. The range of the scaled values can affect the GA performance. High variation in scaled values result in rapid reproduction and prevent the GA from searching other regions in the search space. On the other hand, lower scaled value variations give the same opportunity for reproduction resulting low search space progress. *Figure 1* depicted the proposed K-means cluster based GA selection scheme framework.

GKA$_{opt}$

It is quite clear that the validation of clusters plays a vital role to improve the performance of cluster based GA. Validity of clusters is done to measure the quality of the clustering methods based on the compactness and separateness of the clusters. There are two major types of approaches for clusters validation:

External index is used to measure the extent to which cluster labels match externally supplied class labels, e.g. Rand and Adjusted Rand index.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

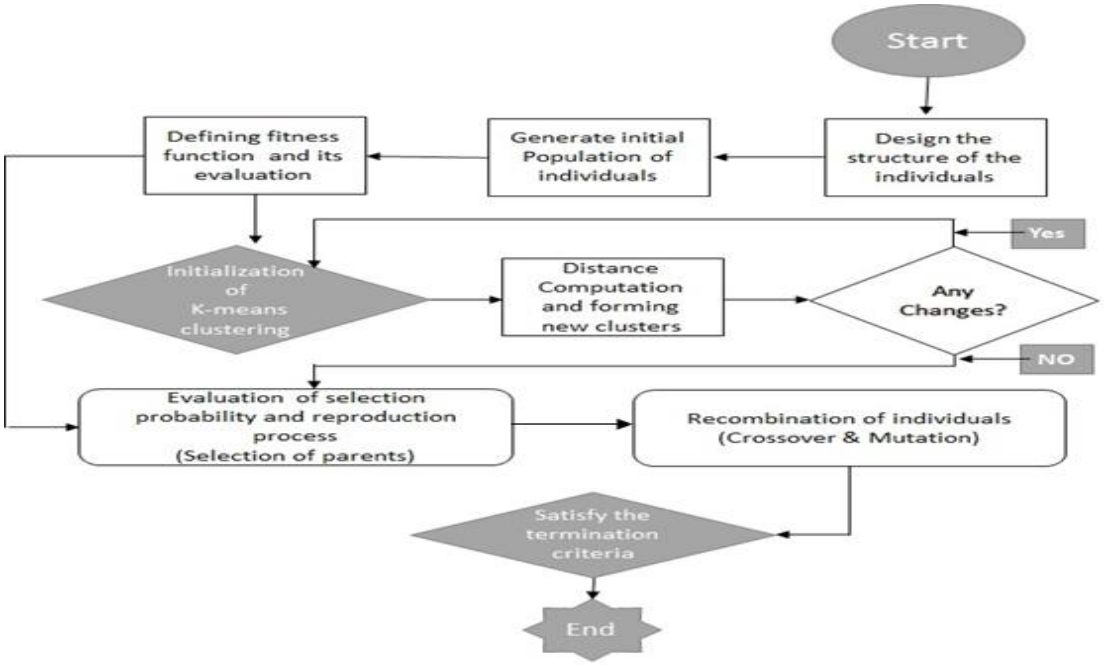

Figure 1. *Frame work of proposed cluster based selection process*

Internal index is used to measure the goodness of a clustering structure without respect to external information e.g. Davies-Bouldin index, Dunn's index, Xie-Beni index, Silhouette index.

The Silhouette index (Mahi, et al., 2018) is one of many cluster indices that can be used to find the number of clusters. The highest value of the average of the Silhouette index s_i indicates a suitable number of clusters. This index is an internal cluster index that can be used as a tool to find the suitable numbers of K with a graphical aid to show the performance of the clustering algorithm.

For each datum i, the Silhouette index can be defined as follows in *Equations 8* and *9:*

$$s_i = \frac{b_{(i)} - a_{(i)}}{max\,(a_{(i)}, b_{(i)})} \qquad \text{(Eq.8)}$$

which is

$$s_i = \begin{cases} 1 - a_{(i)}/b_{(i)}, & \text{if } a_{(i)} < b_{(i)} \\ 0, & \text{if } a_{(i)} = b_{(i)} \\ b_{(i)}/a_{(i)} - 1, & \text{if } a_{(i)} > b_{(i)} \end{cases} \qquad \text{(Eq.9)}$$

where $a_{(i)}$ the average dissimilarities of i with all data points in the same cluster and $b_{(i)}$ the average dissimilarity of i between the other neighbouring cluster. The smallest values of $a_{(i)}$ will indicate the better cluster while the largest values of $b_{(i)}$ will represent a cluster badly matched to its neighbour. If the value of (s_i) is close to 1, then we can say that it is well-clustered. However, if (s_i) is near to 0, then it is not clear in which cluster i belongs to. The larger the average value of the Silhouette index (s_i), the better the performance of the results.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

GKA is a single objective optimization method that use only one fitness function. DB index (Mahi, et al., 2018) is used as their fitness function to measure the validity of the clustering algorithm. This index measured the similarity between the clusters (how separated and compact are the clusters). The lowest values of the index indicate the better clustering and how well the clusters are separated. For these reasons, this study uses the DB index.

The DB index measures the similarity of the clusters by calculating the function of the ratio of the sum within cluster scatter to the between cluster separation. The scatter within the C_i for the i^{th} can be computed as in *Equation 10:*

$$S_{i,q} = \left(\frac{1}{|C_i|}\sum_{x \in C_i} \|x - c_i\|_2^q\right)^{1/q} \quad \text{(Eq.10)}$$

where c_i is the center of the cluster C_i. Usually the choice of q is *2*, where a Euclidean distance measured between the center of the cluster and the individual data points. The $R_{i,qt}$ in *Equation 11* denotes the similarity of C_i to the other clusters. In this study, the Hamming distance is denoted in $d_{ij,t} = d(C_i;C_j)$.

$$R_{i,qt} = \mathrm{Max}_{j,j \neq i}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}\right\} \quad \text{(Eq.11)}$$

where $d_{ij,t} = d(C_i, C_j) = \|c_i - c_j\|t$.

Hence the fitness value can be evaluated by:

$$DB_i = \frac{1}{k_i}\sum_{i=1}^{k_i} R_{i,qt} \quad \text{(Eq.12)}$$

From *Equation 10,* the measure of dispersion of a cluster C_i; $i = 1,\ldots\ldots,K_i$ is represented by $S_{i,q}$. The evaluation of the DB index for the chromosome Ch_i is defined in *Equation 12* where the lowest value is indicated the better clustering. The whole process of GKA_{opt} technique focuses on finding optimum number of clusters within a range of clusters through evaluation of validity index functions. Hence the number of clusters K and the searching for k^{th} is constrained to a suitable interval $[K_{min}; K_{max}]$, where $K_{min} > 2$ and $K_{max} \leq n = 10$. The individual has a fixed length of K_{max}. While, the number of K is held to fixed then $K_{min} = K_{max}$. The reason to hold the K constant is to make sure that the best number of clusters has been pre-specified and its validity tested by the Silhouette index, whereas the minimum value of DB determines the optimization of clusters.

Benchmark functions

There are several optimization procedures claiming dominance over other procedures. Hence, to obtain an optimum solution, benchmark functions can be utilized as indicators to authenticate their effectiveness. Many benchmark functions along with their properties have been used to appraise the feasibility of optimization problems. Hence the efficiency of algorithm is based on the nature of the optimization problem regarding variation rate in objective function and the number of local optima etc. (Zhang et al., 2013). A multimodal benchmark function has a minimum of two local optima. The efficient search technique effectively eliminates the region around local optima for searching global optima. Furthermore, the dimensionality of search space is

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

another significant factor which makes the problem more complicated. A comprehensive study regarding dimensionality problem and its characteristics were carried out by Friedman. During the search process, the value regarding global optimum needs to be obtained efficiently. Hence the areas close to local minima must be avoided as possible. If the local optima are randomly distributed in the search area, then that is considered to be a more difficult problem. The optimization process is focused on obtaining the global optimum point, consequently the regions nearby local optima should be circumvented because the optimization process might be stuck at local optima and then the local optima is considered as global optima (Deng et al., 2015).

To evaluate the performance and sustainability of the proposed selection operators, we will use eleven unimodal, multi-modal, non-separable, convex and continuous benchmark functions. *Table 2* presents the list of benchmark functions (Surjanovic and Bingham, 2016) utilized to appraise the efficiency of suggested evolutionary methods. Hence the benchmark function name, limit, properties and its fitness function are presented in *Table 2*. These benchmark functions have varying complexities that are most commonly applied in many comparative studies. The necessary detail regarding these benchmarks are given below:

Table 2. *Detail of benchmark functions for comparison*

Benchmark	Fitness function	Search limits	Optimum value	Properties
Axis parallel ellipsoid	$f(x) = \sum_{i=1}^{D} ix_i^2$	[-5.12, 5.12]	0	Continuous, convex, unimodal
Bohachevsky	$f(x) = \sum_{i=1}^{D} (x_i^2 + 2x_{i+1}^2 + 0.3 - 0.3\cos(3\pi x_i + 4\pi x_{i+1}))$	[-100, 100]	0	Multimodal, non-separable
Booth	$f(x) = \sum_{i=1}^{D} ((x_i + 2x_{i+1} - 7)^2 + (2x_i + x_{i+1} - 5)^2)$	[-10, 10]	0	Multimodal, separable
Easom	$f(x) = \sum_{i=1}^{D} \cos(x_i)\cos(x_{i+1})e^{(-(x_i-\pi)^2-(x_{i+1}-\pi)^2)}$	[-100, 100]	-1	Multimodal, non-separable
Ellipsoidal	$f(x) = \sum_{i=1}^{D} (100^{\frac{i-1}{k-1}}x_i)^2$	[-5.12, 5.12]	0	Multimodal, non-separable
Himmelblau	$f(x) = \sum_{i=1}^{D} ((x_i^2 + x_{i+1} - 11)^2 + (x_i + x_{i+1}^2 - 7)^2)$	[-6, 6]	0	Non-convex, multimodal
Six-hump camel	$f(x) = \sum_{i=1}^{D} \{(4 - 2.1x_i^2 + \frac{1}{3}x_i^4)x_i^2 + x_i x_{i+1} + 1 + (-4 + 4x_i^2 + 1)x_i^2\}$	[-3, 3]	(-0.0898, -0.7126, 0.0898, 0.7126)	Multimodal, non-separable
Matyas	$f(x) = \sum_{i=1}^{D} (0.26(x_i^2 + x_{i+1}^2) - 0.48x_i x_{i+1})$	[-10, 10]	0	Non-scalable, Unimodal
Maccormick	$f(x) = \sum_{i=1}^{D} (\sin(x_i + x_{i+1}) + (x_i + x_{i+1})^2 - 1.5x_i + 2.5x_{i+1} + 1)$	[-3,-1.5, 4,4]	-1.91333	Multimodal, Non-separable
Rastrigin	$f(x) = \sum_{i=1}^{D} [10n + \{(x_i^2 - x_{i+1})^2 + (1 - x_i)^2\}]$	[-5.12, 5.12]	0	Multimodal, Separable
Schwefel	$f(x) = \sum_{i=1}^{D} x_i \sin(\sqrt{\lvert x_i \rvert})$	[-500, 500]	0	Multimodal, Non-separable

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Statistical results and discussion

The statistical results of GKA methods (**GKA$_{opt}$** (S-index), **GKA$_{opt}$** (DB-inbex), **GKA$_{P1}$** and **GKA$_{P2}$**) were evaluated at 10, 50 and 100 dimensions and compared with standard GA. In the present experimental study, the performance of optimization techniques is evaluated by fixed parameters such as population size, maximum number of generations, crossover fraction and scaling function. The population size for each experiment is 50 along with 0.8 two-point crossover fraction for 10, 50 and 100 dimensions. Each experiment is executed thirty times to determine the statistical results in terms of means, standard deviation (S.D) and t-test. An independent t-test is obtained to assess the significant difference between standard GA and proposed selection techniques. The performance of these selection methods are evaluated on eleven benchmark functions using MATLAB version R2015a.

The statistical results in *Tables 3–5* reveal that the proposed selection techniques (**GKA$_{P1}$** and **GKA$_{P2}$**) outperform under all benchmark functions from 10 to 100 dimensions. Additionally, the probability values of Bohachevsky, Easom and Schwefel benchmark functions at dimension 10 are ranging 0.0010 to 0.0093 which is significant to some extent but when we increase the dimension of the experiment from 50 to 100, the experimental results turned into highly significant with probability values are tend toward 0.0000. Most of the statistical results demonstrate that **GKA$_{P2}$** perform the best for achieving an optimal solution and **GKA$_{P1}$** is the second best technique because of its closeness to the theoretical optimum value. Overall cluster based selection techniques are broader and more comprehensive for achieving optimum solution and also restrict the individuals to premature convergence. More specifically, there are slight differences between optimum values of **GKA$_{P1}$** and **GKA$_{P2}$** at lower dimension but these differences become highly significant at 50 to 100 dimensions. Results of the above tables are also reveal that the proposed selection techniques perform distinctly better in unimodal and multimodal benchmark functions but rate of the change is slightly high in unimodal functions which turn into highly significant results.

In the context of the above results it is described that the cluster based GA techniques establish a realistic partitioning of the population. These techniques actually extend the population diversity by strengthening the search process and limiting the chance of less fitted individuals. The inclusion of the cluster is also more beneficial for reducing selection pressure, hence the proposed selection schemes authenticate the process of best fitted individuals' selection. Additionally, partitioning of the individuals in the form of clusters endorse that the adequate mixture of individuals is always carried forward to the next generation for obtaining optimum solutions. Above statistical results ensure that clustering is always helpful in exploration process and also minimizing the chance of permute convergence at local optima due to well-adjusted selection pressure.

In order to evaluate the pairwise comparison between the above selection techniques: a non-parametric statistical test is used known as Wilcoxon matched pair signed rank test. The test statistic (T_c) of this test is based on the ranking of absolute difference between two techniques. T^+ is the rank sum with positive signs and T^- is the rank sum with negative signs, the value of Tc depends on the fewer rank sum between T^+ and T^-. A sufficiently small value of T^+ and T^- will cause the rejection of the null hypothesis. The results in *Table 5* represent the pairwise comparison of the following cluster based selection techniques using Wilcoxon matched pair signed rank test at 5% level of significance.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Table 3. *Statistical comparison of the optimum values for different selection techniques under 10 dimensions*

Selection methods (Dimension 10)						
Benchmark	**Statistics**	**GKA$_{opt}$ (S index)**	**GKA$_{opt}$ (DB index)**	**GKA$_{P1}$**	**GKA $_{P2}$**	**Standard GA**
Axis parallel hyper ellipsoid	Mean	1.64×10^{-7}	1.93×10^{-7}	2.70×10^{-7}	2.66×10^{-7}	5.08×10^{-5}
	S.D	1.93×10^{-7}	2.03×10^{-7}	2.14×10^{-7}	2.10×10^{-7}	7.98×10^{-7}
	T-test	0.0010	0.0006	0.0000	0.0000	---
Bohachevsky	Mean	0.17	0.19	0.20	0.20	0.54
	S.D	0.41	0.44	0.34	0.34	0.61
	T-test	0.0027	0.0026	0.0012	0.0010	---
Booth	Mean	13.30	14.60	14.80	15.00	17.50
	S.D	1.03×10^{-4}	1.07×10^{-4}	1.08×10^{-4}	1.09×10^{-4}	2.01×10^{-1}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Easom	Mean	-0.90	-0.84	-1.08	-1.07	-4.37
	S.D	0.25	0.17	0.39	0.38	1.47
	T-test	0.0028	0.0031	0.0085	0.0079	---
Ellipsoidal	Mean	3.28×10^{-3}	3.37×10^{-3}	3.71×10^{-3}	3.64×10^{-3}	3.56×10^{1}
	S.D	0.01	0.01	0.01	0.01	123.00
	T-test	0.00083	0.00081	0.00053	0.00029	---
Himmelblau	Mean	-1.55×10^{4}	-1.98×10^{4}	-2.25×10^{4}	-2.28×10^{4}	-1.63×10^{8}
	S.D	4.89×10^{3}	5.12×10^{3}	8.06×10^{3}	8.20×10^{3}	2.06×10^{7}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Six-Hump	Mean	-1.03	1.27	-1.70	-1.67	-2.05
	S.D	0.08	0.11	0.14	0.14	0.15
	T-test	0.0000	0.0000	0.0000	0.0000	---
Matyas	Mean	3.21×10^{-3}	2.51×10^{-3}	5.30×10^{-3}	5.21×10^{-3}	3.75×10^{-2}
	S.D	4.89×10^{-3}	5.29×10^{-3}	5.24×10^{-3}	5.15×10^{-3}	5.28×10^{-2}
	T-test	0.0000	0.0000	0.0008	0.0008	---
Maccormick	Mean	-3.81	-4.45	-5.29	-5.38	-6.13
	S.D	0.38	0.44	0.49	0.48	1.71
	T-test	0.0000	0.0000	0.0008	0.0000	---
Rastrigin	Mean	7.90×10^{2}	7.38×10^{2}	9.03×10^{-2}	8.87×10^{-2}	9.90×10^{-2}
	S.D	1.38×10^{-4}	1.07×10^{-4}	1.67×10^{-4}	1.65×10^{-4}	3.33×10^{-1}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Schwefel	Mean	-2.06×10^{3}	-2.41×10^{3}	-2.73×10^{3}	-2.68×10^{3}	-4.02×10^{3}
	S.D	2.15×10^{2}	2.28×10^{2}	2.39×10^{2}	2.35×10^{2}	1.38×10^{2}
	T-test	0.0053	0.0046	0.0029	0.0025	---

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Table 4. *Statistical comparison of the optimum values for different selection techniques under 50 dimensions*

Selection methods (Dimension 50)						
Benchmark	**Statistics**	**GKA$_{opt}$ *(S index)***	**GKA$_{opt}$ *(DB index)***	**GKA$_{P1}$**	**GKA $_{P2}$**	**Standard GA**
Axis parallel hyper ellipsoid	Mean	15.40	13.50	16.00	15.70	315.00
	S.D	7.90	8.12	8.39	8.24	185.00
	T-test	0.0000	0.0000	0.0000	0.0000	---
Bohachevsky	Mean	11.70	11.10	12.50	12.20	78.20
	S.D	2.10	2.02	2.35	2.30	34.00
	T-test	0.0000	0.0000	0.0000	0.0000	---
Booth	Mean	93.10	93.05	91.70	93.50	153.00
	S.D	3.36	3.17	3.45	3.51	25.50
	T-test	0.0000	0.0000	0.0000	0.0000	---
Easom	Mean	-1.25	-1.02	-1.49	-1.47	-10.20
	S.D	0.28	0.19	0.52	0.51	4.08
	T-test	0.0000	0.0000	0.0000	0.0000	---
Ellipsoidal	Mean	0.63×10^{4}	0.56×10^{4}	1.01×10^{4}	0.87×10^{3}	8.18×10^{4}
	S.D	1.02×10^{4}	0.88×10^{4}	1.04×10^{4}	1.02×10^{4}	5.35×10^{4}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Himmelblau	Mean	-1.07×10^{9}	-1.22×10^{9}	-7.05×10^{4}	-7.18×10^{4}	-1.45×10^{9}
	S.D	6.11×10^{7}	6.43×10^{7}	2.50×10^{4}	2.55×10^{4}	1.42×10^{8}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Six-Hump	Mean	-2.79	-3.17	-3.32	-3.38	-7.35
	S.D	0.34	0.38	0.41	0.42	2.79
	T-test	0.0000	0.0000	0.0000	0.0000	---
Matyas	Mean	0.56	0.79	0.90	0.88	2.60
	S.D	0.22	0.28	0.35	0.34	0.81
	T-test	0.0000	0.0000	0.0000	0.0000	---
Maccormick	Mean	35.40	35.60	37.30	36.60	59.90
	S.D	1.73	1.75	1.80	1.76	4.78
	T-test	0.0000	0.0000	0.0000	0.0000	---
Rastrigin	Mean	4.37×10^{3}	$4.42\times\times10^{3}$	4.43×10^{3}	4.52×10^{3}	4.97×10^{3}
	S.D	1.08	1.96	1.98	2.02	2.83
	T-test	0.0000	0.0000	0.0000	0.0000	---
Schwefel	Mean	-6.47×10^{4}	-5.36×10^{4}	-7.92×10^{3}	-7.77×10^{3}	-11.54×10^{4}
	S.D	7.22×10^{2}	6.99×10^{2}	7.39×10^{2}	7.25×10^{2}	8.49×10^{2}
	T-test	0.0000	0.0000	0.0000	0.0000	---

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Table 5. *Statistical comparison of the optimum values for different selection techniques under 100 dimensions*

Selection methods (Dimension 100)						
Benchmark	**Statistics**	**GKA$_{opt}$ (S index)**	**GKA$_{opt}$ (DB index)**	**GKA$_{P1}$**	**GKA $_{P2}$**	**Standard GA**
Axis parallel hyper ellipsoid	Mean	626.00	645.00	664.00	649.00	3050.00
	S.D	103.00	127.00	159.00	155.00	823.00
	T-test	0.0000	0.0000	0.0000	0.0000	---
Bohachevsky	Mean	62.20	55.70	68.40	66.90	239.00
	S.D	11.30	10.70	12.70	12.40	44.40
	T-test	0.0000	0.0000	0.0000	0.0000	---
Booth	Mean	209.00	225.00	265.00	271.00	531.00
	S.D	23.00	24.10	26.00	26.50	130.00
	T-test	0.0000	0.0000	0.0000	0.0000	---
Easom	Mean	-1.20	-1.10	-1.65	-1.61	-12.90
	S.D	0.54	0.31	0.65	0.64	1.81
	T-test	0.0000	0.0000	0.0000	0.0000	---
Ellipsoidal	Mean	1.36×10^{5}	1.45×10^{5}	1.74×10^{5}	1.70×10^{5}	1.37×10^{5}
	S.D	4.24×10^{4}	4.55×10^{4}	4.83×10^{4}	4.73×10^{4}	7.32×10^{5}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Himmelblau	Mean	-1.06×10^{5}	-1.09×10^{5}	-1.12×10^{5}	-1.10×10^{5}	-2.73×10^{9}
	S.D	1.17×10^{4}	1.23×10^{4}	2.73×10^{4}	2.67×10^{4}	2.54×10^{8}
	T-test	0.0000	0.0000	0.0000	0.0000	---
Six-Hump	Mean	3.66	3.29	3.83	3.74	47.10
	S.D	3.38	2.79	4.10	4.01	20.00
	T-test	0.0000	0.0000	0.0000	0.0000	---
Matyas	Mean	3.67	3.40	3.77	3.69	10.90
	S.D	0.58	0.43	0.61	0.60	2.55
	T-test	0.0000	0.0000	0.0000	0.0000	---
Maccormick	Mean	86.00	81.00	84.60	86.50	179.00
	S.D	5.83	4.34	5.80	5.93	27.70
	T-test	0.0000	0.0000	0.0000	0.0000	---
Rastrigin	Mean	8.34×10^{3}	8.13×10^{3}	8.57×10^{3}	8.76×10^{3}	9.95×10^{3}
	S.D	2.17	2.14	2.52	2.58	3.22
	T-test	0.0000	0.0000	0.0000	0.0000	---
Schwefel	Mean	-0.97×10^{4}	-0.93×10^{4}	-1.04×10^{4}	-1.02×10^{4}	-2.41×10^{4}
	S.D	1.19×10^{3}	1.23×10^{3}	1.36×10^{3}	1.33×10^{3}	1.71×10^{3}
	T-test	0.0000	0.0000	0.0000	0.0000	---

According to the results indicated in *Table 6*, there is highly significant difference between most of the pairs of K-means cluster based techniques, when the dimension of the experiment increases from 10 to 100 the p-value turn into highly significant. However, there is also a non-significant difference between GKA$_{P1}$ and GKA$_{P2}$ at lower dimension but when dimension increases difference between them becomes highly significant.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Table 6. *Pairwise comparison of different selection techniques using Wilcoxon signed rank test*

Comparison	Dimension	T^+	T^-	T_c	P-value
GKA$_{opt}$(S-index) vs **GKA$_{opt}$**(DB-inbex)	10	112	353	112	0.0160
	50	103	362	103	0.0063
	100	87	378	87	0.0010
GKA$_{P1}$ vs **GKA$_{P2}$**	10	187	278	187	0.1799
	50	163	302	163	0.0790
	100	142	323	142	0.0318
GKA$_{opt}$(S-index) vs **GKA$_{P1}$**	10	127	338	127	0.0147
	50	87	378	87	0.0010
	100	45	420	45	0.0000
GKA$_{opt}$(DB-inbex) vs **GKA$_{P1}$**	10	96	369	96	0.0020
	50	45	420	45	0.0000
	100	28	437	28	0.0000
GKA$_{opt}$(S-index) vs **GKA$_{P2}$**	10	119	346	119	0.0093
	50	78	387	78	0.0005
	100	39	426	39	0.0000
GKA$_{opt}$(DB-inbex) vs **GKA$_{P2}$**	10	82	383	82	0.0007
	50	36	429	36	0.0000
	100	24	441	24	0.0000
GA vs **GKA$_{P1}$ & GKA$_{P2}$**	10	465	0	0	0.0000
	50	465	0	0	0.0000
	100	465	0	0	0.0000

Conclusions and future work

K-means cluster based genetic algorithm (GKA) selection techniques are proposed to solve optimization problems using unimodal and multimodal benchmark functions. Two distinct types of GKA techniques were proposed, one is using fixed number of clusters GKA_F and other is through optimum number of clusters GKA_{opt}. Hence, Davies-Bouldin and Silhouette index is used to determine optimum number of clusters in the population. The GKA method comprise of four major stages: clustering, membership probability indexing, fitness evaluation and selection. The hybridization of genetic algorithm and clustering of individuals will effectively cater the problem of population diversity and selection pressure. A selection probability is assigned to each individual after the process of K-means clustering. Fitness scaling changes the membership fitness scores in a limit that is suitable for selection function, which select the parents for future generation by the utilization of scaled fitness values. The comparative performance of each cluster based GA technique (GKA_{opt} (S-index), GKA_{opt} (DB-inbex), GKA_{P1} and GKA_{P2} are evaluated on eleven benchmark functions under 10, 50 and 100 dimensions. Usually, the performance of standard GA is good to solve unimodal problem but unable to handle multimodal problems. By the hybridization of GA and K-means clustering, the newly proposed selection techniques efficiently and effectively handle the multimodal problems by obtaining optimum value. The statistical results of present study represent that the performance of GKA_{P1} and GKA_{P2} selection operators are comparatively more superior to standard GA selection techniques. In addition, the

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

significance of proposed techniques is also improved by increasing the dimension of the experiment from 50 to 100. In-fact the cluster based selection techniques outperform in solving unimodal and multimodal problems with positive impact.

In the present research study, integrating the strengths of data mining and evolutionary computation were limited to single-objective optimization problems. Future research study could evaluate the performance of GKA in solving constrained optimization problems with multi-objective functions. The efficacy of novel selection procedure should be considered in future research study. It would be compelling to hybridize the K-means cluster with population based optimization schemes such as firefly algorithm (Rajan and Malakar, 2015), ant colony optimization (Gao et al., 2016) and particle swarm optimization (Pednekar, 2019). Finally, another potential avenue for future research is to examine the performance by making comparison of standard K-means clustering with the other clustering methods like Incremental K-means (IKM), Scalable K-means and Online K-means (Saharan et al., 2018) by considering presently proposed selection techniques.

Acknowledgments. Authors are very grateful to deanship of scientific research at King Khalid University, Abha, Saudi Arabia for the financial support through General Research Program under project number GRP-32-41.

Data availability. The data used to support the findings of this manuscript are taken from the website https://www.sfu.ca/~ ssurjano/optimization.html.

Conflict of interests. All authors of this article declare that there is no conflict of interests regarding the publication of this article.

REFERENCES

[1] Aibinu, A. M., Salau, H. B., Rahman, N. A., Nwohu, M. N., Akachukwu, C. M. (2016): A novel clustering based genetic algorithm for route optimization. – Engineering Science and Technology, an International Journal 19(4): 2022-2034.

[2] Belahbib, F. Z. B., Souami, F. (2011): Genetic algorithm clustering for color image quantization. – 3rd European Workshop on Visual Information Processing, IEEE, July 4-6, Paris, pp. 83-87.

[3] Belhaouari, S. B., Ahmed, S., Mansour, S. (2014): Optimized K-means algorithm. – Mathematical Problems in Engineering 2014(2): 1-14.

[4] Das, S., Abraham, A., Konar, A. (2008): Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. – Pattern Recognition Letters 29(5): 688-699.

[5] Deng, Y., Liu, Y., Zhou, D. (2015): An improved genetic algorithm with initial population strategy for symmetric TSP. – Mathematical Problems in Engineering. http://dx.doi.org/10.1155/2015/212794.

[6] Devooght, R. (2010): Multi-objective genetic algorithm. – https://pdfs.semanticscholar.org/bfee/f4cf230d4db51ef332237dae8530f2b5f613.pdf.

[7] Eiben, A. E., Smit, S. K. (2011): Parameter tuning for configuring and analyzing evolutionary algorithms. – Swarm and Evolutionary Computation 1(1): 19-31.

[8] Fister Jr, I., Yang, X. S., Fister, I., Brest, J., Fister, D. (2013): A brief review of nature-inspired algorithms for optimization. – arXiv preprint arXiv 1307.4186.

[9] Gao, S., Wang, Y., Cheng, J., Inazumi, Y., Tang, Z. (2016): Ant colony optimization with clustering for solving the dynamic location routing problem. – Applied Mathematics and Computation 285: 149-173.

[10] Haq, E., Hussain, A., Ahmad, I., IbrahimM, Almanjahie (2019a): A Novel Selection Approach for Genetic Algorithms for Global Optimization of Multimodal Continuous

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

Functions. – Computational Intelligence and Neuroscience. https://doi.org/10.1155/2019/8640218

[11] Haq, E., Hussain, A., Ahmad, I. (2019b): Development a New Crossover Scheme for Traveling Salesman Problem by aid of Genetic Algorithm. – International Journal of Intelligent Systems and Applications (IJISA) 12(2): 46-52.

[12] Hatamlou, A., Abdullah, S., Nezamabadi-Pour, H. (2012): A combined approach for clustering based on K-means and gravitational search algorithms. – Swarm and Evolutionary Computation 6: 47-52.

[13] Huang, H., Ma, Y. (2019): A hybrid clustering approach for bag-of-words image categorization. – Mathematical Problems in Engineering. https://doi.org/10.1155/2019/4275720.

[14] Hussain, A., Muhammad, Y. S., Nauman Sajid, M., Hussain, I., Mohamd Shoukry, A., Gani, S. (2017): Genetic algorithm for traveling salesman problem with modified cycle crossover operator. – Computational Intelligence and Neuroscience. https://doi.org/10.1155/2017/7430125.

[15] Hussain, A., Muhammad, Y. S. (2019): Trade-off between exploration and exploitation with genetic algorithm using a novel selection operator. – Complex & Intelligent Systems. https://doi.org/10.1007/s40747-019-0102-7.

[16] Islam, M. Z., Estivill-Castro, V., Rahman, M. A., Bossomaier, T. (2018): Combining k-means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. – Expert Systems with Applications 91: 402-417.

[17] Jain, A. K. (2010): Data clustering: 50 years beyond K-means. – Pattern Recognition Letters 31(8): 651-666.

[18] Li, Z. Y., Yi, J. H., Wang, G. G. (2015): A new swarm intelligence approach for clustering based on krill herd with elitism strategy. – Algorithms 8(4): 951-964.

[19] Lin, H. J., Yang, F. W., Kao, Y. T. (2005): An efficient GA-based clustering technique. – Journal of Tamkang University of Science and Technology 8(2): 113-122.

[20] Loai, L., Lin, T., Li, B. (2008): Mri brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach. – Pattern Recognition Letters 29(10): 1580-1588.

[21] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., Brown, S. J. (2004): Incremental genetic K-means algorithm and its application in gene expression data analysis. – BMC Bioinformatics 5(1): 172.

[22] Mahi, H., Farhi, N., Labed, K., Benhamed, D. (2018): The silhouette index and the K-harmonic means algorithm for multispectral satellite images clustering. – 2018 International Conference on Applied Smart Systems (ICASS), IEEE, Medea University, Médéa, Algeria, 24-25 November, pp. 1-6.

[23] Maulik, U., Bandyopadhyay, S., Mukhopadhyay, A. (2011): Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics. – Springer Science & Business Media, Berlin.

[24] Murty, M., Babaria, R., Chiranjib, B. (2008): Clustering Based on Genetic Algorithms. – In: Ghosh, A. et al. (eds.) Multi-Objective Evolutionary Algorithms for Knowledge Discovery from Databases. Vol. 98 of Studies in Computational Intelligence. Springer, Berlin, pp. 137-159.

[25] Pednekar, A. M. (2019): Optimal initialization of K-means using Particle Swarm Optimization. – arXiv preprint arXiv 1904.09098.

[26] Rahman, A., Islam, Z. (2011): Seed-detective: A novel clustering technique using high quality seed for K-means on categorical and numerical attributes. – Proceedings of the Ninth Australasian Data Mining Conference (Australian Computer Society, Inc.) 121: 211-220.

[27] Rajan, A., Malakar, T. (2015): Optimal reactive power dispatch using hybrid Nelder–Mead simplex based firefly algorithm. – International Journal of Electrical Power & Energy Systems 66: 9-24.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177

[28] Saharan, S., Baragona, R., Nor, M. E., Salleh, R. M., Asrah, N. M. (2018): Clustering for binary data sets by using genetic algorithm-incremental K-means. – Journal of Physics: Conference Series 995(1): 012038.
[29] Sivanandam, S. N., Deepa, S. N. (2008): Genetic Algorithms. – In: Sivanandam, S. N. (ed.) Introduction to Genetic Algorithms. Springer, Berlin, pp. 15-37.
[30] Steinhaus, H. (1956): Sur la division des corps materiels en parties. – Bull. Acad. Polon. Sci. C1 II-IV: 801-804.
[31] Surjanovic, S., Bingham, D. (2016): Virtual library of simulation experiments: test functions and datasets (2013). – https://www.sfu.ca/ssurjano/optimization.html.
[32] Xu, R., Wunsch, D. C. (2005): Survey of clustering algorithms. – IEEE Transactions on Neural Networks 16(3): 645-678.
[33] Yang, F., Sun, T., Zhang, C. (2009): An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. – Expert Systems with Applications 36(6): 9847-9852.
[34] Yang, X. S. (2014): Nature-Inspired Optimization Algorithms. – Elsevier, Amsterdam.
[35] Younus, Z. S., Mohamad, D., Saba, T., Alkawaz, M. H., Rehman, A., Al-Rodhaan, M., Al-Dhelaan, A. (2015): Content-based image retrieval using PSO and k-means clustering algorithm. – Arabian Journal of Geosciences 8(8): 6211-6224.
[36] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., Zeebaree, S. R. (2017: Combination of K-means clustering with genetic algorithm: a review. – International Journal of Applied Engineering Research 12(24): 14238-14245.
[37] Zhang, M. X., Zhang, B., Zheng, Y. J. (2014): Bio-inspired meta-heuristics for emergency transportation problems. – Algorithms 7(1): 15-31.
[38] Zhang, X., Zhang, Y., Hu, Y., Deng, Y., Mahadevan, S. (2013): An adaptive amoeba algorithm for constrained shortest paths. – Expert Systems with Applications 40(18): 7607-7616.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1706_1415914177