Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Reinforced Genetic Algorithm using Clustering based on Statistical Estimation

Taekyoon Park, Yeonsoo Kim, Jong Min Lee *,1

* *School of Chemical and Biological Engineering, Engineering Development Research Center, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea.*

Abstract: Genetic algorithm(GA) has beend widely used to obtain solution in various optimization problems because of its robustness and convergence property. GA algorithm, however, has limitations; the computational time increases sharply as the complexity of the problem increases and the user's arbitrary judgement is involved especially in its termination step. In order to solve these limitations, we suggest a modified reinforced GA using clustering based on the statistical estimation. In this method, the mathematical reliability is gradually increased for each generation to find the solution vector to be finally obtained. The similarity between each solution vector generated by GA is determined and the inefficient repetitive calculation is remarkably reduced. In addition, the statistical reliability of the obtained solution vectors can be calculated to reduce the randomness of the user in the conventional termination step.

Keywords: Genetic algorithms, optimization, parameter estimation, similarity determination, statistical estimation

1. INTRODUCTION

Genetic algorithm(GA) is one of the search techniques which belongs to the evolutionary algorithm of the guided random search technique. It mimics the evolutionary process in nature and guides search algorithms based on the mechanics of biological evolution. Since it has been developed by John Holland in 1970's, the method is considered as an appropriate approach to determining solutions in large and potentially huge spaces.

Despite its versatility, GA has several drawbacks. Since it is an heuristic-based algorithm, it has a randomness depending on the user at the time of termination. In addition, since the solution vectors obtained for each generation are directly input to the objective function, the computation time is proportional to the complexity of the problem.

In order to reduce computational time, many researches have been conducted. Kim et al. (2001) showed that GA with the less fitness evaluation can be made by combining the existing clustering method. Although the theoretical basis was weak, the introduction of clustering showed the potential for improved GA efficiency.

Deb et al. (2002) proposed a fast and elitist multi-objective GA, NSGA-II, and reduced the computational complexity in GA. The proposed method showed the better performance and alleviated three difficulties: computational complexity, non-elitism, and sharing parameter.

Despite the fact that clustering can improve the performance of GA, there are many studies to improve the clustering based on the opposite GA, rather than applying the clustering to the GA due to the heuristic characteristics of the GA. Bandyopadhyay and Maulik (2002) suggested KGA-clustering, a GA-based efficient clustering technique for utilizing the principles of the existing K-means algorithm. GA guided clustering method was proposed by Bezdek et al. (1994) for unsupervised clustering resulting.

In this paper, we propose a reinforced genetic algorithm using clustering method based on statistical estimation to find a solution faster than general GA within the same time. To reduce the randomness of the most commonly used k-means clustering - the user specifies the number of groups - the newly proposed statistical estimation based clustering computes the similarity based on the distance and angle between solution vectors. As a result of the similarity determination, solution vectors belonging to the same group are treated as having the same objective function value, and only a part of solution vectors obtained as a result are evaluated. This can improve the speed of the GA and the number of clustering can be adjusted appropriately by an exploring function over the range of the standard normal distribution.

2. PRELIMINARIES

2.1 Assumptions

In this paper, four assumptions are made. These conditions correspond to the basic prerequisites for using GA.

a. Every solution has a genetic representation.

It is a basic assumption in order to use GA.

[1] Associate Professor, Corresponding author(e-mail:jongmin@snu.ac.kr)

b. The objective function does not diverge for all values in the solution domain.

If the objective function diverges for some elements in the solution domain, it is improper to obtain the optimal solution.

c. It takes a finite amount of time to solve the optimization problem.

In order to compare the efficiency of existing and newly proposed algorithms, the optimization problems to be dealt with must be resolved within a finite time.

d. There is an optimal solution under given constraints.

If the optimal solution is not within a given range, the necessity of using GA disappears.

2.2 Genetic Algorithm

GA can be represented in 5 different steps: Initialization of population, evaluation, selection, crossover and mutation, and termination (Weile and Michielssen, 1997).

(1) Initialization of population

The population size is determined by user considering the nature of the problem. It is usually generated randomly in the entire range of feasible solution.

(2) Evaluation

For the generated population, value of the objective function is evaluated. It can be multi-objective function.

(3) Selection

Based on the evaluation, elite solution population is selected to breed a new generation.

(4) Crossover and Mutation

In order to generate a new generation population from the selected solution, crossover and mutation method is used. This results in new solution set which is different from the previous solution set.

(5) Termination

When a solution satisfying the minimum condition is found, the generation of a specific numerical value is reached, or the user's arbitrarily conditions are satisfied, GA can be terminated.

2.3 Clustering

The proposed clustering among the solution vectors is conducted by using four mathematical concepts.

Reference vector The reference vector, O_i, is a vector that determines the similarity between solution vectors and is defined as the mean value of solution vectors, s_{i-1,j_k}, obtained at each $(i-1)^{th}$ step of crossover and mutation. The initial value is defined as the value that the user inputs.

$$O_i \equiv \frac{\sum s_{i-1,j_k}}{j} \tag{1}$$

Note that the value of k may vary depending on the value of i.

Radius The radius, r_{i,j_k}, is defined as the Euclidean norm between all k solution vectors and the reference vector of the i^{th} generation.

$$r_{i,j_k} \equiv d(O_i, s_{i,j_k}) \tag{2}$$

Adjacent distance The adjacent distance, d_{i,j_k,j_l}, is defined as the Euclidean norm between the given two solution s_{i,j_k} and s_{i,j_l}.

$$d_{i,j_k,j_l} \equiv d(s_{i,j_k}, s_{i,j_l}) \tag{3}$$

Angle The angle between two solution vectors, θ_{i,j_k,j_l}, is defined as the angles after translating them by the reference vector, O_i.

$$\theta_{i,j_k,j_l} \equiv cos^{-1} \frac{(s_{i,j_k} - O_i) \cdot (s_{i,j_l} - O_i)}{\|s_{i,j_k} - O_i\| \|(s_{i,j_l} - O_i)\|} \tag{4}$$

Criteria The criterion for determining the degree of similarity for the three variables discussed above can be set as follows.

(1) Radius

As the radius decreases, the similarity increases with respect to the given reference vector.

(2) Adjacent distance

The closer the distance is, the higher the degree of similarity.

(3) Angle

The smaller the angle between the two vectors is, the higher the similarity is.

Sub-objective function Since the evaluation time required to obtain the solution is likely to increase when the criterion for each generation is examined, a sub-objective function using an elementary function that can be easily calculated is established.

$$J_{i,j_k,j_l} \equiv \frac{(r_{i,j_k} - r_{i,j_l})^2 + d_{i,j_k,j_l}^2}{\|O_i\|^2} \times tan(\theta_{i,j_k,j_l}) \tag{5}$$

The sub-objective function is calculated for ${}_kC_2$ cases generated from k solution vectors in i^{th} generation. If the two solutions are the same, the function value becomes 0, and the penalty becomes larger as the angle formed by the two solution vectors with respect to the reference vector becomes larger.

Drop out policy The following dropout policy can be established based on the calculated sub-objective function.

If $J_{i,j_k,j_l} \leq K$, s_{i,j_k} and s_{i,j_l} are treated as belonging to the same group, and one of them are discarded which has the smaller radius, r_{i,j_k}.

K is determined by a statistical estimation to be described later.

2.4 Statistical Estimation

Statistical estimation is a technique commonly used to estimate population mean from sample values. As the number of samples increases, the reliability increases, but the efficiency decreases. This property can be combined with the trade off in GA's selection and evaluation. The main idea is similar to Kim et al. (2016).

(1) Mean and Variance

The mean and variance can be defined for three variables (distance from the reference vector, distance between adjacent vectors, and angle between adjacent vectors based on the reference vector). This step is necessary to standardize each variable.

$$r_{m,i} \equiv mean(r_{i,j_k}), v_{r,m,i} \equiv var(r_{i,j_k}) \tag{6}$$

$$d_{m,i} \equiv mean(d_{i,j_k,j_l}), v_{d,m,i} \equiv var(d_{i,j_k,j_l}) \tag{7}$$

$$\theta_{m,i} \equiv mean(\theta_{i,j_k,j_l}), v_{\theta,m,i} \equiv var(\theta_{i,j_k,j_l}) \tag{8}$$

(2) Interval generation

Based on the mean and variance, it is possible to set the interval if we use the standardization method. Now we have the virtual interval where the solutions are treated as belonging toe the same group with a specific mathematical confidence, $Z_{confidence}$.

$$U_{variable} = [m_{variable} - \alpha, m_{variable} + \alpha] \tag{9}$$

where $\alpha = Z_{confidence} \times \frac{\sigma(U_{variable})}{\sqrt{n}}$. For example, if we use a 95% reliability, $Z_{confidence}$=1.96. Here n is the number of $U_{variable}$.

(3) Determination of K K is the boundary value for the drop out policy. It defined in various ways, which is defined in the simplest form, minimum value of interval, in this study.

$$K \equiv min(2\alpha_i, i = r, d, \theta) \tag{10}$$

2.5 Profile of $Z_{confidence}$

When a new generation is produced, it is possible to reduce the number of solution vectors extracted efficiently by controlling the value of $Z_{confidence}$. Depending on the system, various profiles of $Z_{confidence}$ can be possible.

Monotone increasing function Monotone increasing function can reduce the number of solutions extracted and the computational time as the value of $Z_{confidence}$ continues to increase. The degree of reduction of $Z_{confidence}$ according to the generation can be appropriately adjusted according to the needs of the user. This function can be used when the degree of similarity between the solutions obtained is high.

Monotone decreasing function When monotone decreasing function is applied, the number of solution vectors extracted increases over the generations. Using these functions for all intervals has the disadvantage of increasing the amount of computation while the scope of observation is increasing. This strategy can be used when the number of solution vector is too small to proceed with the GA.

It can be seen that the profile of $Z_{confidence}$ can be set flexibly based on the similarity between the solutions obtained and the value of the objective function calculated therefrom. In this study, we use the monotone decreasing function of the exponential function type. Fig. 1 shows the profile of $Z_{confidence}$ used in this study.

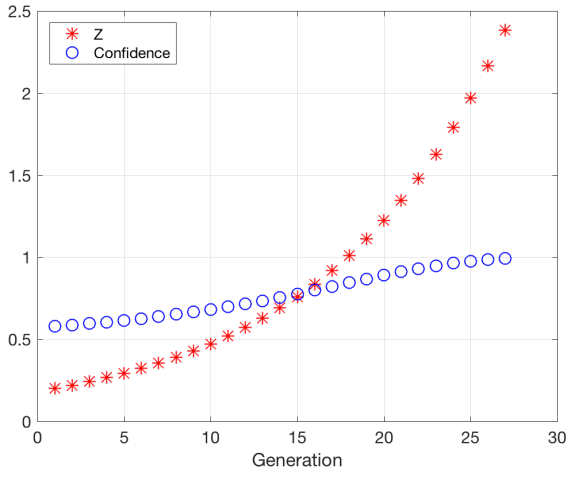

Fig. 1. Profile of $Z_{confidence}$

3. SIMULATION RESULTS AND DISCUSSION

We used the passive selective catalytic reduction(pSCR) system in Kim et al. (2016) as an example to compare the GA with the proposed GA. We predicted the parameters and compare the results with known parameter values. There are three major comparisons: the angle with the actual solution vector, the distance, and the value of the objective function. The objective function is set to the cumulative concentration of NOx and NH_3 that should be minimized in the pSCR. The number of generations was fixed to 35. The results are shown in Fig. 1, 2, and 3. (Red: General GA, Blue: Proposed GA) Under the profile of $Z_{confidence}$ used in this study, the proposed GA did not generate a solution after the 27th generation.

Fig. 2 shows the angle with actual solution vector. Except for the 2nd and 27th generation, the similarity with the reference solution vector is higher than the existing GA in all segments.

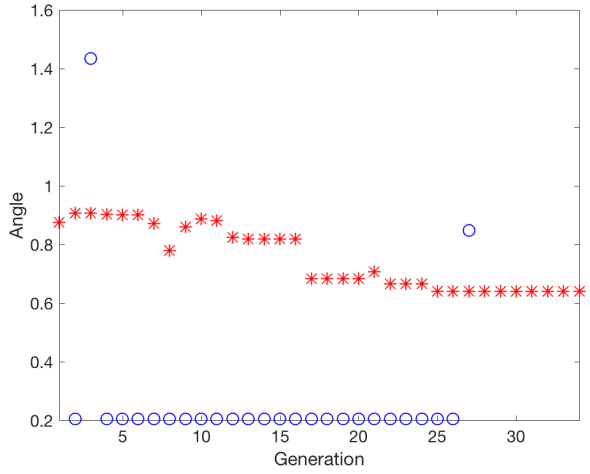

Fig. 2. Angle with actual solution vector

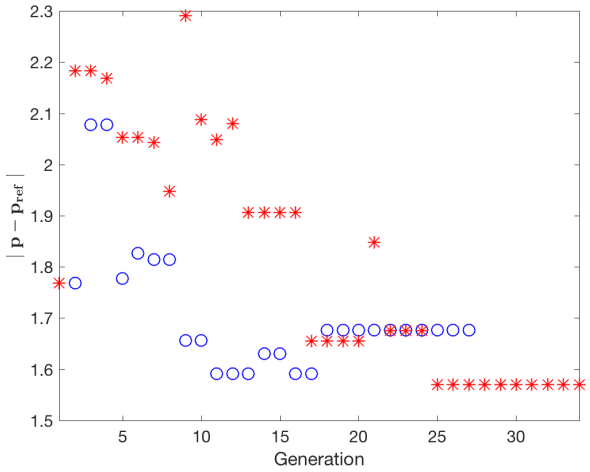

Fig. 3. Distance to the actual solution vector

From 1 to 17 generations, the degree of similarity was higher than that of existing GA, and thereafter, the degree of similarity was lower than that of existing GA. This is shown in Fig. 3.

The objective function showed a tendency to decrease in both the GA and the proposed method, and did not show significant change after 15th generation. After 15th generation, the objective function value of the solution vector obtained from the existing GA corresponds to about half of the value of the proposed objective function. Fig. 4 shows this trend.

Even though the proposed GA showed better performance in angle and distance than the general GA, the calculated value of the objective function was larger than that of the general GA. However, this is due to the fact that the currently set $Z_{confidence}$ profile is a monotone increasing function, and if the profile of Z is controlled, the proposed GA may have an even better result after the 27th generation. In addition, compared to the time spent, the general GA took up to $1.6hr/generation$, but the proposed GA takes $0.57hr/generation$ and showed a speed three times faster than the general GA. (Computation was performed on an Intel i5-4670 3.40 GHz processor.) Therefore, if

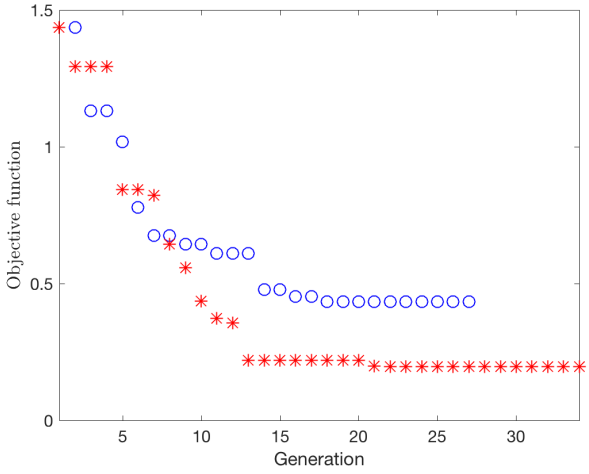

Fig. 4. Value of the objective function

the time required is included in the use of the GA, the proposed GA is sufficiently competitive to general GA. The long computational time of GA makes robustness hard to check. Through the proposed method, it is possible to run GA several times during the same time interval, which can guarantee robustness.

4. CONCLUSION

This paper proposes a reinforced genetic algorithm using clustering based on the statistical estimation. The disadvantage of general GA can be improved by including clustering technique with statistical estimation. K-means clustering method, which is the most popular clustering method, requires the user to specify the number of groups. However, in the proposed method, clustering is performed considering both angle and distance based on the mathematical confidence profile. The statistical estimation is used to reduce the randomness of the GA, increase the calculation speed, and secure the robustness, and it is advantageous to run the GA multiple times within the same time interval. The disadvantage existing in the objective function value can be improved by controlling the profile of $Z_{confidence}$ based on the similarity between the obtained solution vectors and the objective function value. Comparisons with previous studies combining GA and clustering will be made later.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (MSIP) (NRF-2016R1A5A1009592).

REFERENCES

Bandyopadhyay, S. and Maulik, U. (2002) An evolutionary technique based on K-Means algorithm for optimal clustering in RN. *Information Sciences*, 146(1), 221 - 237.

Bezdek, J. C., Boggavarapu, S., Hall, L.O., and Bensaid, A.(1994) Genetic algorithm guided clustering *Proceedings of the First IEEE Conference on Evolutionary*

Computation. IEEE World Congress on Computational Intelligence, 1, 34 - 39.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2), 182 - 197.

Kim, H. and Cho, S. An efficient genetic algorithm with less fitness evaluation by clustering. *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on* 887 - 894.

Kim, Y., Lee, S.J., Park, T., Lee, G., Suh, J.C., and Lee, J.M. (2016) Robust leak detection and its localization using interval estimation for water distribution network. *Computers & Chemical Engineering.*, 92: 117.

Kim, Y., Jung, C., Kim, C.H., Kim, Y.W., and Lee, J.M. (2016). Dynamic modelling and sensitivity analysis integrated LNT-pSCR system. *IFAC-PapersOnLine*, 49(7), 326 - 331. 11th IFAC Symposium on Dynamics and Control of Process SystemsIncluding Biosystems DYCOPS-CAB 2016.

Weile, D.S. and Michielssen, E. (1997) Genetic algorithm optimization applied to electromagnetics: a review. *IEEE Transactions on Antennas and Propagation*, 45(3), 343 - 353.