Development of Discrete-Cockroach Algorithm (DCA) for Feature Selection Optimization

1st Yusuf Hendrawan
Faculty of Agricultural Technology
Universitas Brawijaya
Malang, Indonesia
yusufhendrawan@gmail.com

2nd Muchnuria Rachmawati
Faculty of Agricultural Technology
Universitas Brawijaya
Malang, Indonesia
muchnuria@yahoo.com

3rd Muchammad Riza Fauzy
Graduate School of Technology Management
Institut Teknologi Sepuluh November
Surabaya, Indonesia
Muchrizafauzi@gmail.com

***Abstract*— One of the recently proposed algorithms in the field of bio-inspired algorithm is the Hungry Roach Infestation Optimization (HRIO) algorithm. Haven has developed optimization algorithms HRIO that is inspired by recent discoveries in the social behavior of cockroaches. Result showed that HRIO was effective at finding the global optima of a suite of test functions. However, there is no researcher who has observed HRIO for solving discrete feature selection problems. Therefore, we try to develop a discrete-cockroach algorithm (DCA) as the modification of HRIO for optimizing discrete feature selection problem. We test the algorithm to solve feature selection problem in machine vision for predicting water stress in plant using single and multi-objectives optimization. Two objective functions are minimizing prediction error and minimizing the number of feature-subset. The results showed DCA has better performance compared to the existed bio-inspired optimization algorithms such as genetic algorithms (GA) and discrete-particle swarm optimization (discrete-PSO). The performance showed significant difference between DCA and other methods.**

Keywords*— *discrete cockroach algorithm, feature selection, multi objective optimization

I. Introduction

The natural systems have been one of the rich sources of inspiration for developing new intelligent systems [1; 2]. One of the intelligent system's role is to solve the feature selection problem. In machine learning, discretization and feature selection are important techniques for preprocessing data to improve the performance of an algorithm on high-dimensional data [3]. In this study, a novel artificial intelligence approaches using nature-inspired algorithm for feature selection optimization is proposed. Haven et al [4] have developed optimization algorithms hungry roach infestation optimization (HRIO) that is inspired by recent discoveries in the social behavior of cockroaches. HRIO that has been developed is still used to solve optimization problems in general cases, but there has never been a study that tested the performance of HRIO to solve feature selection issues.

HRIO principle is similar to particle swarm optimization (PSO). There are three factors used by PSO and HRIO, namely inertia, personal influence, and social influence. The difference lies in social influence, where PSO uses global best position while HRIO uses local best position. Obagbuwa et al [5] have developed a dynamic step size adaptation roach infestation optimization (DSARIO) to improve the HRIO swarm stability and enhance local and global search performance. To know the effectiveness of a new optimization algorithm hence required comparative optimization method that has been widely used in research such as genetic algorithm (GA) and PSO. GA is search algorithms based on the mechanics of natural selection and natural genetics [6]. PSO is an evolutionary computation technique inspired in the behavior of bird flocks which was first introduced by Kennedy and Eberhart [7]. There have been many studies using PSO to solve discrete problems called discrete-PSO [8; 9].

However, there is still no research that tests HRIO performance to solve discrete feature selection problems with single objective optimization [10] and multi-objective optimization [11]. Obagbuwa and Adewumi [12] made modification of social cockroach behaviors, called modified roach infestation optimization (MRIO). The equations in HRIO can be modified to solve the discrete feature selection problem by substituting the algorithm for personal influence and social influence factors with the crossover and mutation methods used in GA. The result shows that MRIO can find global optima of multi-dimensional functions. The objective of this study is developing discrete cockroach algorithm (DCA) as bio-inspired algorithm for solving feature selection problem using single and multi-objective optimization. The DCA is compared with the forefront bio-inspired optimization algorithms i.e. GA and discrete-PSO.

II. Research Methods

Based on the research conducted by Jeanson [13] about characteristics and behavior of cockroaches, there are three simple behaviors of cockroach agents which can be defined as:

1. *Find_Darkness*: cockroaches search for the darkest location in the search space as shown in Fig. 1. The level of darkness at a location is directly proportional to the value of the fitness function at that location *F(x)*. Perfectly dark condition means that cockroach agent has reached its maximum condition. While perfectly light condition means that cockroach agent has reached a minimum point.
2. *Find_Friends*: cockroaches enjoy the company of friends and socialize with nearby cockroaches with the probability per unit time ($1/\tau_{stop,N}$) of stopping when encountering N friends: 0.49/s for $N = 1$, 0.63/s for $N = 2$ and 0.65/s for $N = 3$ [13] as shown in Fig. 2. If a cockroach agent comes within a detection radius of another cockroach agent, then there is a probability of $1/\tau_{stop,N}$ that these cockroaches will socialize (or group). This socializing is emulated in the algorithm by a

sharing of information, where this information is the darkest known location of each individual cockroach agent which can be defined as personal best solution (p). In essence, when two cockroach agents meet, there is a chance that they will communicate their knowledge of the search space to each other. They share their knowledge with their neighbors (N) and set the darkest local location of the group which can be defined as local best solution (l).

3. *Find_Food*: when a cockroach agent becomes hungry ($hunger_i$) it searches for food as shown in Fig. 3. The food locations are initialized randomly in the search space. The *Find_Food* behavior periodically perturbs the population, ideally minimizing the chance of converging to local optima.

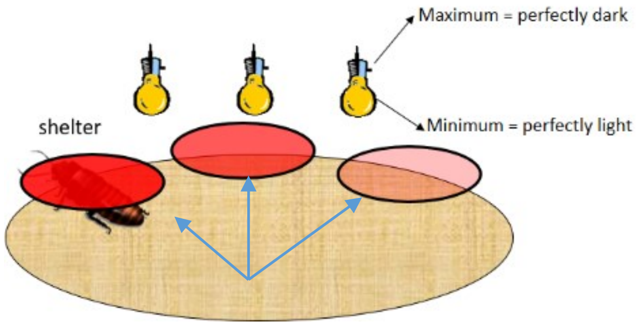

Fig. 1. Find darkness behaviour.

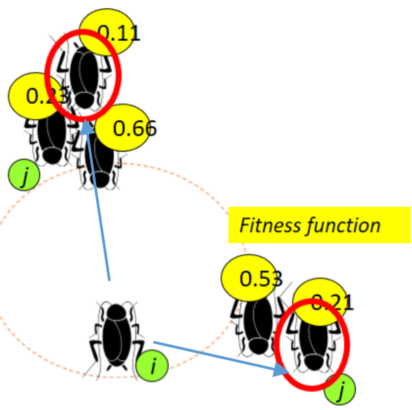

Fig. 2. Find friends behavior.

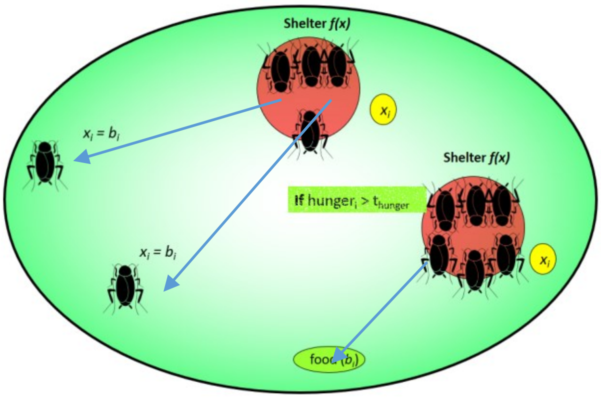

Fig. 3. Find food behavior.

The algorithm of proposed DCA as shown in Fig. 4 are as follows:

1. Initialization of DCA parameters. The maximum iteration (t_{max}) is set by the user. In this case we set the global iteration (t_{max} = 500). The number of cockroach population (N_a) is = 70. For neighbors updating, the parameters are A_1 = 0.49, A_2 = 0.63 and A_3 = 0.65. For hunger updating, t_{hunger} = 100. The probability of mutation is set (w = 0.5). The probability of crossover (C_o) is set in various values.
2. Generate cockroach location (x_i) randomly and $hunger_i = rand\{0, t_{hunger}-1\}$. Each cockroach consist of discrete solution (e.g. x_i: 0,1,1,0,0,0,0,1,0,0,1,0,…..m), where m is the number of problem space. Each x_i in the population represents a candidate solution to the discrete problem.
3. Evaluate each cockroach agent (x_i) using fitness function.
4. Update the individual solution $F(x_i)$. Individual solution $F(x_i)$ is calculated according to the fitness value (x_i).
5. Calculate neighbors' threshold value (d_g):

$$M = [M_{jk}] = \frac{\|F(x_j) - F(x_k)\|}{2} \quad (1)$$

$$d_g = median\{M_{jk} \in M : 1 \le j < k \le N_a\} \quad (2)$$

6. Repeat steps 6.1 to 6.4 for those x_i with partial solutions. Steps 6.1 to 6.8 are as follows:

6.1 Updating personal best solution (p_i) for the individual cockroach agent. For minimizing objective function:

$$p_i \begin{cases} p_i = x_i & if\ F(x_i) < F(p_i) \\ p_i & otherwise \end{cases} \quad (3)$$

For maximizing objective function:

$$p_i \begin{cases} p_i = x_i & if\ F(x_i) > F(p_i) \\ p_i & otherwise \end{cases} \quad (4)$$

6.2 Compute the neighbors (N_i) of cockroach i.
For k = 1 to N_a

$$N_i \begin{cases} N_i = N_i + 1 & if\ k : 1 \le k \le N_a, k \ne i\ AND\ M_{ik} < d_g \\ N_i & otherwise \end{cases} \quad (5)$$

6.3 Update the darkest local location or group best solution (l_i) according to:
For r = 1 to N_i

$$l_i \begin{cases} l_i = l_{j_r} = \arg\min_k \{F(p_k)\}, k = \{i, j_r\} & if\ rand[0,1] < A_{\min\{N_i,3\}} \\ l_i & otherwise \end{cases} \quad (6)$$

where $\{i, j\}$ are the indices of the two socializing cockroaches and p_k is the darkest known location for the individual cockroach agent personal best.

6.4 Update roach location (x_i):

$$x_i \begin{cases} x_i = C_o \oplus CR(C_o \oplus CR(w \oplus MT(x_i), p_i), l_i) & if\ hunger_i < t_{hunger} \\ x_i = random & otherwise \end{cases} \quad (7)$$

The update x_i consists of three components: The first component is $a_i = w \oplus MT(x_i)$, which represents the velocity of the cockroach. MT represents the mutation operator with the mutation probability of w. In other words, a uniform random number rnd[0, 1] is generated. If rnd[0, 1] is less than w then single insert move mutation operator is applied. The second component is $b_i = C_o \oplus CR(a_i, p_i)$, which is the cognition part of the cockroach agent representing the private thinking of the cockroach agent itself. CR represents the crossover operator between a_i and p_i with the probability of C_o. Two points crossover (point1 and point2) are selected randomly, where point1<point2, point1>1 and point2<m. The third component is $x_i = C_o \oplus CR(b_i, l_i)$, which is the social part of the cockroach agent representing the collaboration among the group. CR represents the crossover operator between b_i and l_i with the probability of C_o.

6.5 Evaluate each feature-subset (x_i) using fitness function.

6.6 Update the individual solution $F(x_i)$ based on fitness value.

6.7 Update $hunger_i$:

$$hunger_i = hunger_i + rand[0,1] * t_{hunger} \quad (8)$$

6.8 Update iteration-best solution T^{IB}.

For minimizing objective function:

$$T^{IB} = \arg\min q(F(x_i)) \quad (9)$$

For maximizing objective function:

$$T^{IB} = \arg\max q(F(x_i)) \quad (10)$$

where function $q(.)$ gives the quality of the solution.

7. Update the total best solution T^{TB} by the current iteration-best solution T^{IB}.

For maximizing objective function:

$$T^{TB}\begin{cases} T^{TB} & \text{if } q(T^{TB}) \geq q(T^{IB}) \\ T^{IB} & \text{otherwise} \end{cases} \quad (11)$$

For minimizing objective function:

$$T^{TB}\begin{cases} T^{TB} & \text{if } q(T^{TB}) \leq q(T^{IB}) \\ T^{IB} & \text{otherwise} \end{cases} \quad (12)$$

8. Update best feature-subset.
9. Stopping criterion: the algorithm stops with the total-best solution T^{TB} and best feature-subset. The search will terminate if the global iteration has been reached.

Testing is done for optimizing feature selection in machine vision for predicting water stress in plant using artificial neural network modeling. Sunagoke Moss is used as the experimental plant to study water stress changes in plant. Bhurga et al [14] have tested the use of leaf color for drought stress analysis in rice. The research demonstrates the capability of machine vision for stress level prediction. Testing procedures includes: first process is image acquisition in a dark chamber as shown in Fig. 5, in which the plant images were captured using digital camera (Nikon Coolpix SQ, Japan) placed at 330 mm perpendicular to the sample surface. The image size was 1024 x 768 pixels. Imaging was done under controlled and well distributed light conditions. Light was provided by two 22W lamps (EFD25N/22, National Corporation, Japan). Light intensity over the moss surface was uniform at 100 µmol m^{-2} s^{-1} PPF (Photometer, Li6400, USA) during image acquisition. For single objective and multi-objective optimization, total of 212 features are used, consist of color features and textural features from various kind of color spaces e.g. RGB, Lab, Luv, HIS, HSL, etc. Indriani et al [15] have successfully tested the use of colors and textural features in pattern recognition for biological objects. For modeling the relationship between image features and water stress in plant, artificial neural network (ANN) is used with back-propagation neural network (BPNN) as the learning method. Quiros et al [16] has proven the effectiveness of ANN for pattern recognition using machine vision. The performance of prediction accuracy is measured with root mean square error (RMSE). The number of data used in this study was 500 image data consisting of various water stress condition.

Selection process for selecting relevant image features is done using three nature-inspired approaches i.e. DCA, GA, and discrete-PSO. Multi objective optimization concerns optimization problems with multiple objectives. Barocio et al [17] and Qingqi et al [18] have proven the superiority of bio-inspired algorithms in solving multi-objective optimization. The fitness of multi objective optimization is calculated as follows:

$$function_1 = weight_1 \times RMSE_{(x)} \quad (13)$$

$$function_2 = weight_2 \times \frac{IF_{(x)}}{f_t} \quad (14)$$

$$fitness(x) = function_1 + function_2 \quad (15)$$

where $RMSE_{(x)}$ is the Root Mean Square Error of validation-set data of BPNN using only the expression values of the selected image features in a subset x, where $IF_{(x)}$ is the number of selected image features in x. f_t is the total number of image features, $weight_1$ and $weight_2$ are two priority weights corresponding to the importance of the accuracy and the number of selected image features, respectively, where $weight_1 \in [0.1, 0.9]$ and $weight_2 = 1 - weight_1$. In this study, the accuracy is more important than the number of selected image features in a feature-subset. Problems solved for single and multi objectives optimization in this study use the same training data. In single objective optimization, the objective function is determined by minimizing prediction error. Whereas in multi objective optimization two objective functions are determined by minimizing prediction errors and minimizing the number of feature-subset.

III. Results and Discussion

Figure 6 shows the performance of three bio-inspired optimizations to optimize the feature selection problems in machine vision for single objective optimization. Experimental results show competitive performance among all feature selection optimization techniques. It shows the superiority of DCA, since it achieved better optimization performance as the objective of this research i.e. prediction accuracy of water stress in plant. Based on the observation using some various values of crossover rate ranged from 0.1 to 0.9, GA reached the lowest (optimum) fitness value at crossover rate of 0.5. Discrete-PSO reached the lowest (optimum) fitness value at crossover rate of 0.8. The best fitness plots of the iteration of each optimization method are displayed in Fig. 7 to highlight the search process in each optimization technique. From Fig. 6, the best optimization performance based on fitness value is achieved with DCA followed by GA, and discrete-PSO.

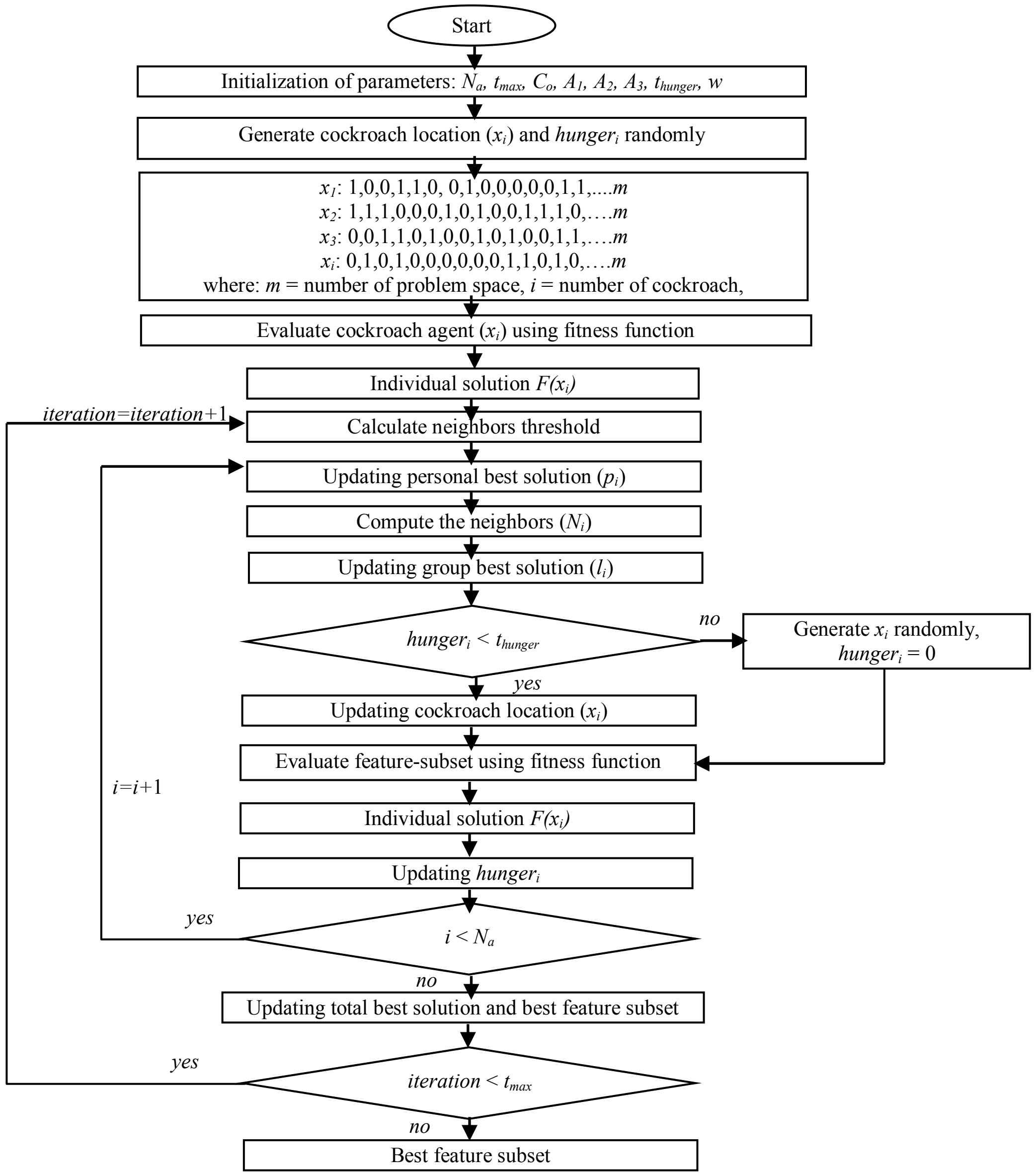

Fig. 4. DCA for discrete-optimization.

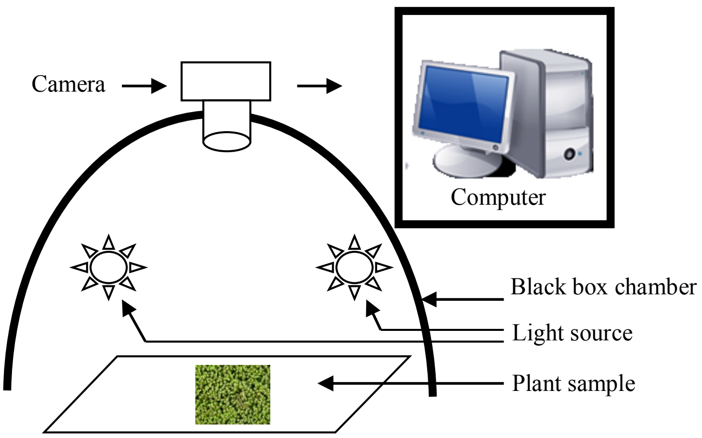

Fig. 5. Procedure of image acquisition and image features extraction

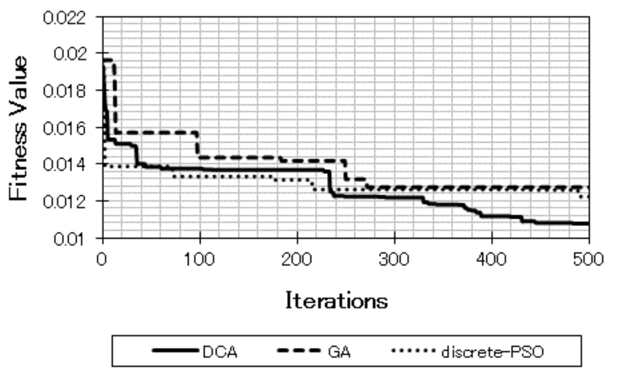

Fig. 6. Performance of single objective discrete-optimization using DCA, GA, and discrete-PSO.

Based on partial analysis using analysis of *t-test* it was shown that there is a significant statistical difference between DCA and other optimization methods at α = 0.05 significant level. Analysis of *t-test* between DCA and GA showed a value of 0.0197, while the *t-test* value between DCA and discrete-PSO showed a value of 0.0143. DCA also showed the least absolute deviation, followed by discrete-PSO, and GA in that order. This implies that DCA showed the highest reliability in optimization process. The results presented in Fig. 6 and 7 clearly show that the DCA have a positive effect on its ability to find global optima for single objective feature selection optimization.

DCA has also been tested using multi-objectives optimization problem [9]. The plots of best fitness values of multi-objectives optimization using all optimization methods are displayed in Fig. 8 to highlight the search process in each optimization method. At the beginning of the iteration, all optimization methods (DCA, GA, and discrete-PSO) were given the same optimization problem which is defined as the initial condition. The fitness value obtained from the initial condition and then normalized by the value of 1.00. During the multi-objectives optimization process the fitness value continues to decrease, searching for the most minimum fitness value. Using the same *weight* parameter ($weight_1$ = 0.9 and $weight_2$ = 0.1), it shows that DCA has the best performance to minimize the fitness value (normalized fitness value = 0.66), followed by GA (normalized fitness value = 0.67), and discrete-PSO (normalized fitness value = 0.84) in that order, respectively. Figure 9 shows the performance of the discrete optimization method with the first objective function which is to minimize the prediction error (RMSE) of water stress in plant.

For the first objective i.e. minimizing prediction error, based on partial analysis using analysis of *t-test* it was shown that there is a significant statistical difference between DCA and GA method at α = 0.05 significant level, but there is no significant difference between DCA and discrete-PSO. Analysis of *t-test* between DCA and GA showed a value of 0.0382, while the *t-test* value between DCA and discrete-PSO showed a value of 0.2567. Figure 10 shows the performance of the discrete optimization method with the second objective function which is to minimize the number of feature-subset. For the second objective i.e. minimizing number of feature-subset, based on partial analysis using analysis of *t-test* it was shown that there is a significant statistical difference between DCA and other optimization methods at α = 0.01 significant level. Analysis of *t-test* between DCA and GA showed a value of 0.000017, while the *t-test* value between DCA and discrete-PSO showed a value of 0.000162.

Most of all optimization methods can quickly minimize the fitness value at the beginning of 50 iterations, but based on the comparison analysis on the performance of all optimization methods, it shows the superiority of DCA to minimize the fitness value in early iterations, followed by GA and discrete-PSO, respectively. From the results, we can see that, DCA is quicker in locating the optimal solution. DCA has the ability to converge quickly. It has strong search capability in the problem space and can efficiently find optimum solution for multi-objectives optimization. From the results obtained, we can see the advantages of DCA in solving single objective optimization problems and multi-objective optimization compared to the other two methods, namely GA and discrete-PSO. This is because DCA has the characteristics to move randomly during the *find food behavior* process so it does not get stuck on local optima. This random *find food behavior* is not owned by GA or discrete-PSO. In addition, DCA has a *find friends behavior* process that is neither owned by GA nor discrete-PSO. In this process global optima will be achieved by relying on best solution from its nearest neighbor and not from the best solution of the swarm population as done by discrete-PSO. This can prevent optimization on local optima.

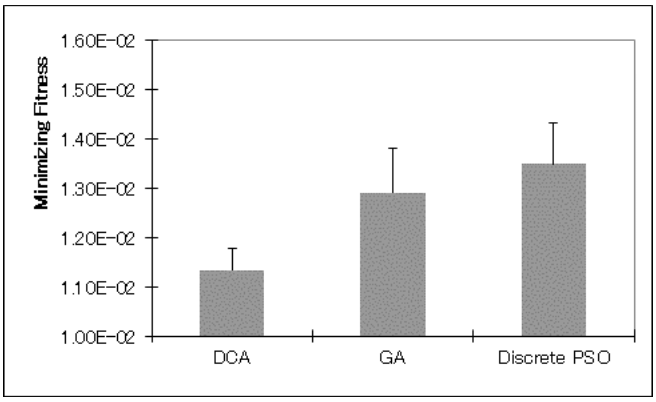

Fig. 7. Comparison of the performance of bio-inspired algorithms for single objective discrete optimization.

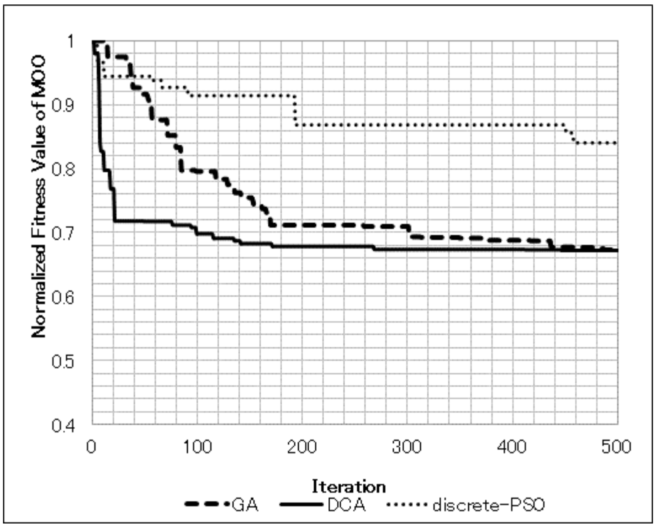

Fig. 8. Performance of multi objectives discrete-optimization (a) DCA; (b) GA; (c) discrete-PSO.

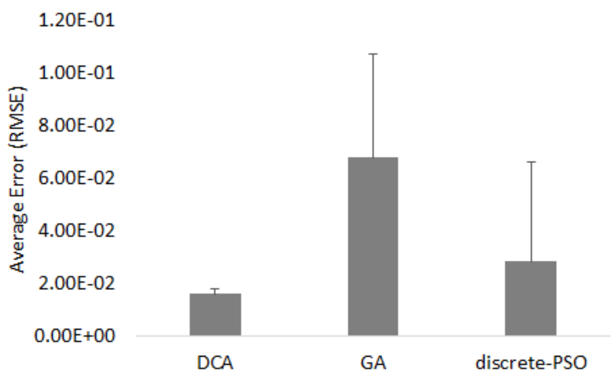

Fig. 9. Performance of multi objectives discrete-optimization for minimizing prediction Root Mean Square Error (RMSE).

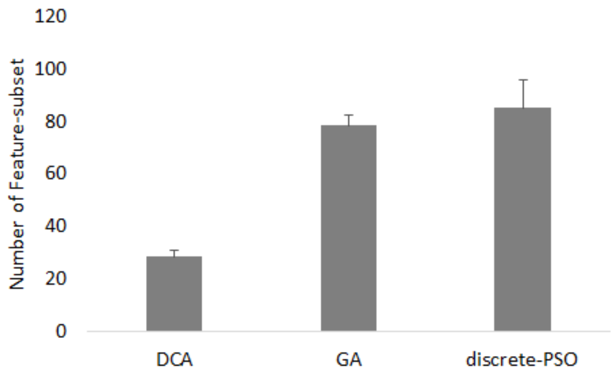

Fig. 10. Performance of multi objectives discrete-optimization for minimizing number of feature-subset.

IV. CONCLUSIONS

Discrete cockroach algorithm (DCA) was developed to search optimum solution with the single objective function and multi-objectives feature selection optimization. To test the performance of DCA as optimization method we compare it with two forefront bio-inspired optimization methods i.e. genetic algorithm (GA), discrete particle swarm optimization (discrete-PSO). The achieved optimization results are promising. DCA has the best performance for solving feature selection optimization problems in machine vision for predicting water stress in plant using single objective optimization and also multi-objectives optimization. There is averagely significant difference of feature selection performance between DCA and other optimization methods. DCA has powerful exploration ability to find global optima. Several characteristics make DCA a unique approach: (1) *Find_Food* behavior that discourages the swarm from converging to local optima; (2) *Find_Friends* behavior to find global best solution which is determined by a group best solution from the neighborhood of the agents to avoid converging on a sub-optimal solution.

REFERENCES

[1] Hendrawan, Y. and Murase, H., "Neural-Genetic Algorithm as feature selection technique for determining sunagoke moss water content", Engineering in Agriculture, Environment and Food (EAEF); Asia Agricultural and Biological Engineering Association (AABEA) Journal, 3(1), pp25-31, 2010

[2] Hendrawan, Y. and Murase, H., "Non-destructive sensing for determining Sunagoke moss water content: Bio-inspired approaches", Agricultural Engineering International: CIGR Journal. Manuscript No. 1564, Vol. 13, No.1, 2011

[3] Tran, B., Xue, B., Zhang, M., "A new representation in PSO for discretization-based feature selection", IEEE Transaction on Cybernetics, 48(6), pp1733-1746, 2018

[4] Havens, T. C., Spain, C. J., Salmon, N. G., Keller, M, "Roach Infestation Optimization", IEEE Swarm Intelligence Symposium. St. Louis MO USA, September 21-23, 2008

[5] Obagbuwa, I.C., Adewumi, A.O., Adebiyi, A.A., "A dynamic step-size adaptation roach infestation optimization", IEEE International Advance Computing Conference (IACC), Gurgaon, India, February 21-22, 2014

[6] Goldberg, D. E., "Genetic algorithms in search, optimization and machine learning", Addison Wesley Longman, Inc, United States of America, 1989

[7] Kennedy, J. and Eberhart, R. C., "Particle swarm optimization", Proceedings of the IEEE International Conference on Neural Networks. Vol 4, pp: 1942-1948, ISBN: 0780327683, Perth, Western Australia November 1995, IEEE, 1995

[8] Hendrawan, Y. and Murase, H., "Neural-intelligent water drops algorithm to select relevant textural features for developing precision irrigation system using machine vision", Comput. Electron. Agr, 77(2), pp214-228, 2011

[9] Hendrawan, Y. and Murase, H., "Bio-inspired feature selection to select informative image features for determining water content of cultured Sunagoke moss", Expert Syst. Appl, 38(11), pp14321-14335, 2011

[10] Hendrawan, Y. and Murase, H., "Neural-discrete hungry roach infestation optimization to select informative textural features for determining water content of cultured Sunagoke moss", Environ. Control Biol, 49(1), pp1-21, 2011

[11] Hendrawan, Y. and Al Riza, D.F, "Machine vision optimization using Nature –inspired algorithms to model Sunagoke moss water status", International Journal on Advance Science Engineering Information Technology, 6, pp2088-5334, 2016

[12] Obagbuwa, I.C., and Adewumi, A.O., "Modified roach infestation optimization", IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Honolulu, HI, USA, May21-24, 2014

[13] Jeanson, R., Rivault, C., Deneubourg, J.L., Blanco, S., Fourniers, R., Jost, C., Theraulaz, G, "Self-organized aggregation in cockroaches", J. Anbehav, 69, pp169-180, 2005

[14] Bhurga, S., Chaudhury, S., Lall, B., "Use of leaf colour for drought stress analysis in rice", IEEE Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Patna, India, December 16-19, 2015

[15] Indriani, O.R., Kusuma, E.J., Sari, C.A., Rachmawanto, E.H., Setiadi, D.R.I.M, "Tomatoes classification using K-NN based on GLCM and HSV color space", IEEE International Conference on Innovative and Creative Information Technology, Salatiga, Indonesia, November 2-4, 2017

[16] Quiros, A.R.F., Abad, A., Bedruz, R.A., Uy, A.C., Dadios, E.P., "A genetic algorithm and artificial neural network-based approach for the machine vision of plate segmentation and character recognition", IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, Cebu City, Philippines, December 9-12, 2015

[17] Barocio, E., Ragalado, J., Cuevas, E., Uribe, F., Zuniga, P., Torres, P.J., "Modified bio-inspired optimization algorithm with a centroid decision making approach for solving a multi-objective optimal power flow problem", IET Generation, Transmission & Distribution, 11(4), pp1012-1022, 2017

[18] Qingqi, Z., Yanling, Q., Yue, L., Nantian, Q.,"Multi-objective mapping for network-on-chip based on bio-inspired optimization algorithms", IEEE Prognostics and System Health Management Conference, Zhangiiaijie, China, August 24-27, 2014