Clustering using Flower Pollination Algorithm and Calinski-Harabasz Index

Szymon Łukasik*†, Piotr A. Kowalski*†, Małgorzata Charytanowicz†‡ and Piotr Kulczycki*†
*Faculty of Physics and Applied Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: {slukasik,pkowal,kulpi}@agh.edu.pl
†Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {slukasik,pakowal,mchmat,kulpi}@ibspan.waw.pl
‡ Institute of Mathematics and Computer Science
The John Paul II Catholic University of Lublin
Konstantynów 1 H, 20-708 Lublin, Poland
Email: mchmat@kul.lublin.pl

***Abstract*—Task of clustering, that is data division into homogeneous groups represents one of the elementary problems of contemporary data mining. Cluster analysis can be approached through variety of methods based on statistical inference or heuristic techniques. Recently algorithms employing novel metaheuristics are of special interest – as they can effectively tackle the problem under consideration which is known to be NP-hard. The paper studies the application of nature-inspired Flower Pollination Algorithm for clustering with internal measure of Calinski-Harabasz index being used as optimization criterion. Along with algorithm's description its performance is being evaluated over a set of benchmark instances and compared with the one of well-known K-means procedure. It is concluded that the application of introduced technique brings very promising outcomes. The discussion of obtained results is followed by areas of possible improvements and plans for further research.**

I. Introduction

CLUSTERING or cluster analysis in computer science refers to the data analysis problem of finding groups in data, in a way that objects belonging to one group are similar to each other and at the same time, there is a significant dissimilarity between different groups' elements. Historically the concept of partitioning a set of objects into disjoint groups was introduced by Aristotle and Theophrastus, but the term "cluster analysis" appeared for the first time around 1930 – in the fields of anthropology and psychology [1]. Currently clustering, as data mining problem, manifests itself in numerous disciplines of science and engineering, e.g. in automatic control [2], text analysis [3], agriculture [4] or marketing [5]. Variety of methods have been also established to tackle this problem. For an overview of these techniques one could refer to [6] or [7].

Clustering, as an optimization task, is known to represent a NP-hard complexity [8], with numerous heuristic approaches being used to find representative groups. Among those iterative partitioning approach of K-means [9] is the most popular one. It tries to minimize the within-cluster sum of squares (WCSS), however it is known to converge to a local solution, without a guarantee of reaching the global optimum. More recent approaches involve using metaheuristic techniques (or their K-means hybrids) to alleviate this issue [10], [11], [12].

This contribution introduces a clustering procedure based on Flower Pollination Algorithm (FPA) proposed in 2012 by Xin-She Yang [13]. It is an example of optimization algorithm mimicking social mechanisms identified in nature [14] – which recently find a variety of applications [15]. As an optimization criterion a value of Calinski-Harabasz index [16] – one of the internal validity measures commonly used for evaluating clustering solution – is proposed.

The paper is organized as follows. First the general description of the Flower Pollination Algorithm, as well as Calinski-Harabasz index, is provided. Then methodological aspects of its application in clustering are being discussed. The results of experimental evaluation along with comparative analysis are covered in the subsequent part of the paper. Finally general remarks regarding algorithms's features and planned further studies are under consideration.

II. Methodological Preliminaries

A. Flower Pollination Algorithm

FPA is an iterative population-based nature-inspired optimization technique aimed at tackling problems of continuous optimization. Solving this class of problems is equivalent to finding x^* which satisfies:

$$f(x^*) = \min_{x \in S} f(x), \tag{1}$$

where $S \subset R^D$, and $f(x)$ constitutes solution's x cost function value. Therefore actual task of the optimizer is to find argument minimizing f.

978-1-5090-0623-6/16/$31.00

To address the optimization task (1) by means of population-based metaheuristic the group of P individual agents is used. It is represented by a set of D-dimensional vectors – equivalent to individuals' positions – within the iteration k denoted by:

$$x_1(k), x_2(k), ..., x_P(k). \tag{2}$$

Euclidean distance between two swarm members, indexed p_1 and p_2 is denoted here by $d(x_{p_1}, x_{p_2})$.

The best position found by given individual p prior to iteration k is given by $x_p^*(k)$ with cost/fitness function value $f(x_p^*(k))$. At the same time:

$$x^*(k) = \arg \min_{p=1,...,P} f(x_p(k)), \tag{3}$$

or

$$x^*(k) = \arg \max_{p=1,...,P} f(x_p(k)), \tag{4}$$

corresponds to the best solution found by the algorithm in its k iterations, with $f(x^*(k))$ representing its related cost (3) or fitness (4) function value. The formula used depends on type of optimization task (minimization – as introduced in (1) – or maximization of function f) [17].

Flower Pollination Algorithm tries to mimic a set of complex mechanisms crucial to the success of plants reproductive strategies in the optimization domain. A single flower or pollen gamete constitutes a solution of the optimization problem, with the whole flower population being actually used. Their constancy will be understood as solution fitness. Pollen will be transferred in the course of two operations used interchangeably, that is: global and local pollination. The first one employs pollinators to carry pollen to long distances towards individual characterized by higher fitness. Local pollination on the other hand occurs within limited range of individual flower thanks to pollination mediators like wind or water [17].

Flower Pollination Algorithm's formal description using aforementioned notation and concepts will be given below (as Algorithm 1). It can be seen that global pollination occurs with probability $prob$ defined by so called switch probability. If this phase is omitted local pollination takes place instead. The first one constitutes of pollinator's movement towards best solution $x^*(k)$ found by the algorithm, with s representing D-dimensional step vector following a Lévy distribution :

$$L(s) \sim \frac{\lambda \Gamma(\lambda) \sin(\pi\lambda/2)}{\pi s^{1+\lambda}}, (s \gg s_0 > 0), \tag{5}$$

with Γ being the standard gamma function and parameters $\lambda = 1.5, s_0 = 0.1$ as suggested by Yang [18]. Practical method for obtaining step sizes s following this distribution by means of Mantegna algorithm are given in [19]. Local pollination includes two randomly selected members of the population and is performed via movement towards them, with randomly selected step size ϵ. Finally, the algorithm is terminated when number of iteration k reaches predetermined limit defined by K [17].

Algorithm 1 Flower Pollination Algorithm [17]

$k \leftarrow 1$ {initialization}
$f(x^*(0)) \leftarrow \infty$
for $p = 1$ to P **do**
  Generate_Solution($x_p(k)$)
end for
{find best}
for $p = 1$ to P **do**
  $f(x_p(k)) \leftarrow$ Evaluate_quality($x_p(k)$)
  if $f(x_p(k)) < f(x^*(k-1))$ **then**
    $x^*(k) \leftarrow x_p(k)$
  else
    $x^*(k) \leftarrow x^*(k-1)$
  end if
end for
{main loop}
repeat
  for $p = 1$ to P **do**
    if $Real\_Rand\_in\_(0,1) < prob$ **then**
      {Global pollination}
      $s \leftarrow Levy(s_0, \gamma)$
      $x_{trial} \leftarrow x_p(k) + s(x^*(k) - x_p(k))$
    else
      {Local pollination}
      $\epsilon \leftarrow Real\_Rand\_in\_(0,1)$
      $r, q \leftarrow Integer\_Rand\_in(1, M)$
      $x_{trial} \leftarrow x_p(k) + \epsilon(x_q(k) - x_r(k))$
    end if
    {Check if new solution better}
    $f(x_{trial}) \leftarrow$ Evaluate_quality(x_{trial})
    if $f(x_{trial}) < f(x_p(k))$ **then**
      $x_p(k) \leftarrow x_{trial}$
      $f(x_p(k)) \leftarrow f(x_{trial})$
    end if
  end for
  {find best and copy population}
  for $p = 1$ to P **do**
    if $f(x_p(k)) < f(x^*(k-1))$ **then**
      $x^*(k) \leftarrow x_p(k)$
    else
      $x^*(p) \leftarrow x^*(k-1)$
    end if
    $f(x(k+1)) \leftarrow f(x_k)$
    $x(k+1) \leftarrow x(k)$
  end for
  $f(x^*(k+1)) \leftarrow f(x^*k)$
  $x^*(k+1) \leftarrow x^*(k)$
  $stop\_condition \leftarrow Check\_stop\_condition()$
  $k \leftarrow k + 1$
until $stop\_condition$ = **false**
return $f(x^*(k))$, $x^*(k)$, k

B. Clustering and its Validation with Calinski-Harabasz Index

Let Y to denote $M \times N$ data matrix:

$$Y = \left[\begin{array}{cccc} y_1 & y_2 & ... & y_M \end{array} \right]^T. \quad (6)$$

Each of N columns of this matrix will be referred later in this paper as a feature and one of M rows as a dataset element or a case.

The task of clustering is equivalent to finding an assignment of data elements $y_1, ..., y_M$ to one of sets (clusters) $CL_1, CL_2, ..., CL_C$. For each non-empty cluster CL_i it is useful to define its centroid $u_i \in R^N$ using:

$$u_i = \frac{1}{M_i} \sum_{y_j \in CL_i} y_j, \quad i = 1, ..., C \quad (7)$$

where M_i constitutes cluster's i cardinality. Similarly U represents the center of gravity of the whole dataset:

$$U = \frac{1}{M} \sum_{j=1}^{M} y_j. \quad (8)$$

Finding adequate partition matching dataset's structure remains a challenging task, even for known number of clusters C. The procedure of estimating how well a clustering recovers natural groups present in the dataset is known as a cluster validation [20]. If correct solution is unavailable solely internal validation techniques, i.e. using only partitioned data, can be used. Calinski-Harabasz index belongs to this group of methods. It is expressed as a ratio of between-cluster variance and the overall within-cluster variance:

$$I_{CH} = \frac{N - C}{C - 1} \frac{\sum_{i=1}^{C} d(u_i, U)}{\sum_{i=1}^{C} \sum_{x_j \in CL_i} d(x_j, u_i)} \quad (9)$$

Well-defined clustering solutions yield high values of I_{CH} index. In a recent comparative study this index was demonstrated to be one of the best cluster validation tools [21], it is therefore used in this paper as a crucial element of algorithm's cost function.

III. Using FPA in Clustering

For solving clustering problem with heuristic optimization algorithm two important aspects need to be settled beforehand, namely: solution representation and cost function definition.

In the approach introduced here clustering solution is represented by a vector of cluster centers:

$$x_p = [u_1, u_2, ..., u_C]. \quad (10)$$

It means that the effective dimensionality D of solution vector x_p is equal to $C * N$.

To evaluate the clustering each data element y_i is assigned to the cluster with the closest center. Then cost function $f(x_p)$ for individual p is obtained using formula:

$$f(x_p) = \frac{1}{I_{CH,p}} + \#_{CL_{i,p}=\emptyset,\, i=1,...,C} \quad (11)$$

that is by adding to the inverse value of Calinski-Harabasz index – calculated for solution p – the number of empty clusters found in this clustering solution denoted by $\#_{CL_{i,p}=\emptyset,\, i=1,...,C}$. It essentially penalizes partitions which do not contain desired number of clusters.

IV. Experimental Studies

One of the main goals of conducted experiments was to examine the relative performance of FPA-based procedure. Popular K-means algorithm was used as a point of reference. The following part of the paper covers the details of algorithms' computational evaluation.

For computational experiments a set of standard synthetic clustering benchmark instances known as S-sets was used [22]. Figure 1 demonstrates structure of those two dimensional data sets. It can be seen that they are characterized by different degree of cluster overlapping.

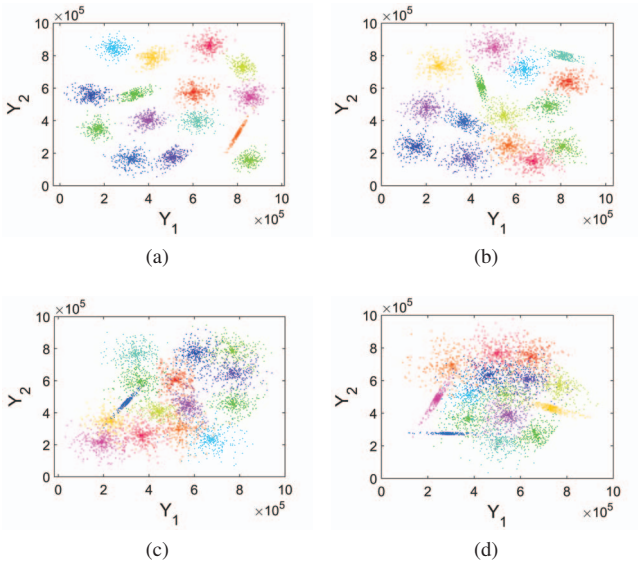


Fig. 1: Scatter plots for *s1* (a), *s2* (b), *s3* (c) and *s4* (d) datasets

To make the study more representative six additional real-world problems, taken from UCI Machine Learning Repository were taken into consideration [23]. Table I characterizes all datasets used in this paper for experimental evaluation. For each instance it reports dataset size along with number of clusters C which can be identified in its structure. For non-synthetic problems it was naturally assumed that each class manifests itself as a single cluster.

As a performance indicator Rand index [24] which measures similarity between clusterings was used. Solutions obtained with investigated algorithms were compared with the reference cluster/class labels. The Rand index is characterized by a value between 0 and 1, with 0 suggesting that the two clusterings do not agree on any pair of points and 1 indicating that they represent exactly the same solution.

K-means clustering and FPA-based clustering were executed 30 times. For FPA we used a population of $P = 20$ individuals, the algorithm was terminated when $C * N * 1000$ cost function evaluations were achieved. It naturally made the

TABLE I: Experimental datasets description

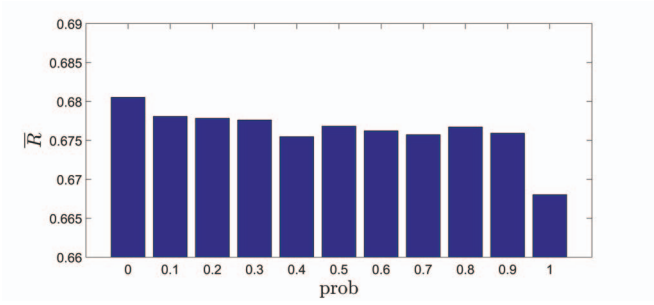
Dataset	*M*	*N*	*C*
s1	5000	2	15
s2	5000	2	15
s3	5000	2	15
s4	5000	2	15
glass	214	9	6
wine	178	13	3
iris	150	4	3
seeds	210	7	3
heart	270	13	2
yeast	1484	8	10

length of search space exploration process dependent on its dimensionality. Summary of obtained results was provided in Table II. It lists mean values of Rand index $\overline{R}$ along with its standard deviation $\sigma(R)$. Among studied algorithms FPA-based clustering was found to be the better-performing one. It not only achieves high performance but is also less prone to entrapment by local minima. Relation between algorithm's performance indicators was also studied by means of pairwise T-tests. In majority of cases the advantage of FPA clustering proved to be statistically meaningful at 0.99 significance level.

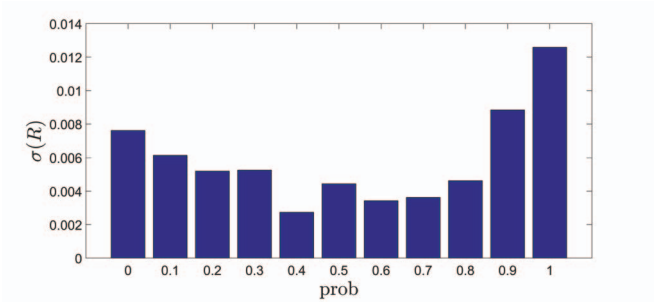
TABLE II: Clustering accuracy measured with Rand index

	K-means clustering		FPA clustering		
	$\overline{R}$	$\sigma(R)$	$\overline{R}$	$\sigma(R)$	Stat. signif.
s1	0.980	0.009	0.995	0.002	Yes
s2	0.974	0.010	0.984	0.003	Yes
s3	0.954	0.006	0.959	0.003	Yes
s4	0.944	0.006	0.949	0.002	Yes
glass	0.619	0.061	0.677	0.005	Yes
wine	0.711	0.014	0.730	0.000	Yes
iris	0.882	0.029	0.893	0.000	Yes
seeds	0.877	0.027	0.884	0.000	No
heart	0.522	<0.001	0.523	0.000	Yes
yeast	0.679	0.041	0.735	0.008	Yes

Presented results where obtained for recommended switch probability value $prob = 0.8$ established in our previous experiments [17]. It is a single adjustable coefficient present in the algorithm scheme. To demonstrate the effect of this parameter we have performed additional experiments with varying $prob$ for the *glass* dataset (the one for which the value of $\sigma(R)$ was the highest). Figure 2 demonstrates means and standard deviations of Rand index obtained in this test. It can be seen that the algorithm is not very sensitive to the alterations of parameter $prob$, with the exclusion of boundary cases (close to 0 and 1). It is another positive feature of the clustering algorithm constructed with the use of Flower Pollination Algorithm.



(a)



(b)

Fig. 2: Mean (a) and standard deviation (b) of Rand index for varying switch probability (*glass* dataset)

V. Conclusion

The paper examined a possibility of using modern nature-inspired technique of Flower Pollination Algorithm for cluster analysis tasks, with Calinski-Harabasz index being employed as the essential element of algorithm's cost function.

We found that FPA-based solution offers high clustering accuracy. For majority of investigated benchmark problems it outperformed standard K-means algorithm both in terms of mean quality of obtained solutions and stability of final results. It means that proposed approach can be used in a variety of real-world engineering tasks were superior accuracy is needed e.g. image segmentation, fuzzy modeling and control [25].

Follow-up studies will include hybridization of FPA with K-means as it can be achieved for any population based optimization technique [26]. Using other well-performing cluster validation indices like Silhouette, COPand and SDbw could be under investigation. As Flower Pollination Algorithm was found to be very effective in multiobjective optimization [27] an option of FPA-based cluster analysis taking into account additional important clustering aspects could be explored as well. Finally, further experiments involving other novel metaheuristic algorithms along with respective comparative analysis are being planned.

REFERENCES

[1] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Berlin-Heidelberg: Springer, 2011.

[2] S. Łukasik, P. Kowalski, M. Charytanowicz, and P. Kulczycki, "Fuzzy models synthesis with kernel-density-based clustering algorithm," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, vol. 3, Oct 2008, pp. 449–453.

[3] C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 77–128. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-3223-4_4

[4] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, "Complete gradient clustering algorithm for features analysis of X-Ray images," in *Information Technologies in Biomedicine*, ser. Advances in Intelligent and Soft Computing, E. Piętka and J. Kawa, Eds. Springer Berlin Heidelberg, 2010, vol. 69, pp. 15–24. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13105-9_2

[5] H. Müller and U. Hamm, "Stability of market segmentation with cluster analysis - a methodological approach," *Food Quality and Preference*, vol. 34, pp. 70 – 78, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950329313002309

[6] L. Rokach and O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 321–352. [Online]. Available: http://dx.doi.org/10.1007/0-387-25465-X_15

[7] C. Aggarwal and C. Reddy, *Data Clustering: Algorithms and Applications*, ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis, 2013. [Online]. Available: https://books.google.pl/books?id=edl7AAAAQBAJ

[8] W. J. Welch, "Algorithmic complexity: three np- hard problems in computational statistics," *Journal of Statistical Computation and Simulation*, vol. 15, no. 1, pp. 17–25, 1982. [Online]. Available: http://dx.doi.org/10.1080/00949658208810560

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66*, 1967, pp. 281–297.

[10] C.-W. Tsai, W.-C. Huang, and M.-C. Chiang, "Recent development of metaheuristics for clustering," in *Mobile, Ubiquitous, and Intelligent Computing*, ser. Lecture Notes in Electrical Engineering, J. J. J. H. Park, H. Adeli, N. Park, and I. Woungang, Eds. Springer Berlin Heidelberg, 2014, vol. 274, pp. 629–636. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40675-1_93

[11] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis," *Applied Soft Computing*, vol. 10, no. 1, pp. 183 – 197, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494609000854

[12] J. Senthilnath, S. Omkar, and V. Mani, "Clustering using firefly algorithm: Performance study," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 164 – 171, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2210650211000265

[13] X.-S. Yang, "Flower pollination algorithm for global optimization," *Lecture Notes in Computer Science*, vol. 7445, pp. 240–249, 2012.

[14] I. Fister Jr., X. Yang, I. Fister, J. Brest, and D. Fister, "A brief review of nature-inspired algorithms for optimization," *CoRR*, vol. abs/1307.4186, 2013. [Online]. Available: http://arxiv.org/abs/1307.4186

[15] X. Yang, "Metaheuristic optimization: Nature-inspired algorithms and applications," in *Artificial Intelligence, Evolutionary Computing and Metaheuristics - In the Footsteps of Alan Turing*, 2013, pp. 405–420. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29694-9_16

[16] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/03610927408827101

[17] S. Łukasik and P. A. Kowalski, "Study of flower pollination algorithm for continuous optimization," in *Intelligent Systems'2014*, ser. Advances in Intelligent Systems and Computing, P. Angelov, K. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, and S. Zadrożny, Eds. Springer International Publishing, 2015, vol. 322, pp. 451–459. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11313-5_40

[18] X.-S. Yang, M. Karamanoglu, and X. He, "Multi-objective flower algorithm for optimization," *Procedia Computer Science*, vol. 18, pp. 861–868, 2013.

[19] X.-S. Yang, *Nature-Inspired Optimization Algorithms*. London: Elsevier, 2014.

[20] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[21] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031200338X

[22] P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," *Pattern Recognition*, vol. 39, no. 5, pp. 761 – 775, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320305003778

[23] "UCI machine learning repository," http://archive.ics.uci.edu/ml/, access 12.04.2016.

[24] H. Parvin, H. Alizadeh, and B. Minati, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846âĂŞ–850, 1971.

[25] P. Kowalski, S. Łukasik, M. Charytanowicz, and P. Kulczycki, "Data-driven fuzzy modeling and control with kernel density based clustering technique," *Polish Journal of Environmental Studies*, vol. 17, no. 4C, pp. 83–87, 2008.

[26] D. van der Merwe and A. Engelbrecht, "Data clustering using particle swarm optimization," in *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, vol. 1, Dec 2003, pp. 215–220 Vol.1.

[27] X.-S. Yang, M. Karamanoglu, and X. He, "Multi-objective flower algorithm for optimization," *Procedia Computer Science*, vol. 18, pp. 861 – 868, 2013, 2013 International Conference on Computational Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050913003943