Automatic Clustering Using a Synergy of Genetic Algorithm and Multi-objective Differential Evolution

Debarati Kundu[1], Kaushik Suresh[1], Sayan Ghosh[1], Swagatam Das[1],
Ajith Abraham[2] and Youakim Badr[2]

[1]Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata, India

[2] National Institute of Applied Sciences of Lyon, INSA-Lyon, Villeurbanne, France
ajith.abraham@ieee.org, *youakim.badr@insa-lyon.fr*

Abstract — This paper applies the Differential Evolution (DE) and Genetic Algorithm (GA) to the task of automatic fuzzy clustering in a Multi-objective Optimization (MO) framework. It compares the performance a hybrid of the GA and DE (GADE) algorithms over the fuzzy clustering problem, where two conflicting fuzzy validity indices are simultaneously optimized. The resultant Pareto optimal set of solutions from each algorithm consists of a number of non-dominated solutions, from which the user can choose the most promising ones according to the problem specifications. A real-coded representation of the search variables, accommodating variable number of cluster centers, is used for GADE. The performance of GADE has also been contrasted to that of two most well-known schemes of MO.

1. Introduction

Optimization-based automatic clustering algorithms greatly rely on a cluster validity function (optimization criterion) the optima of which appear as proxies for the unknown "correct classification" in a previously unhandled dataset [1]. Different formulations of the clustering problem vary in the optimization criterion used. Most existing clustering methods, however, attempt to optimize just one such clustering criterion modeled by a single cluster validity index. This often results into considerable discrepancies observable between the solutions produced by different algorithms on the same data. The single-objective clustering method may prove futile (as judged by means of expert's knowledge) in a context where the criterion employed is inappropriate. In situations where the best solution corresponds to a tradeoff between different conflicting objectives, common sense advocates a multi-objective framework for clustering.

Although there has been a plethora of papers reporting several single-objective evolutionary clustering techniques (a comprehensive survey of which can be found in [1, 2]), very few research works have so far been undertaken towards the application of evolutionary multi-objective optimization algorithms (EMOA) for pattern clustering [3, 4]. A state-of-the-art literature survey indicates that DE has already proved itself as a promising candidate in the field of evolutionary multi-objective optimization (EMO) [5 – 8]. Earlier it has also been successfully applied to single-objective partitional clustering [9 – 11]. The work reported in [3] is based on Deb *et al.*'s celebrated NSGA (Non Dominated Sorting genetic Algorithm)-II [12] and the clustering method described in [4] is based on PESA (Pareto Evolution based Selection) II [13], and both the algorithms are multi-objective variants of Genetic Algorithm (GA). However, the multi-objective variants of DE have not been applied to the general data clustering problems till date, to the best of our knowledge. Since DE, by nature, is a real-coded population-based optimization algorithm, we here

resort to centroid-based representation scheme for the search variables. A MOO algorithm, in general, ends up with a number of Pareto optimal solutions. Here we consider the Xie-Beni index [14] and the Fuzzy C Means (FCM) measure (J_m) [15] as the objective functions. The performance of GADE has also been contrasted with two best-known EMOA-based clustering methods till date. The first of these is MOCK by Handl and Knowles [4] while the second one is based on NSGA II and was used by Bandyopadhyay *et al*. for pixel clustering in remote sensing satellite image data [3]. Here we report the results for ten representative datasets including the microarray Yeast sporulation data [16].

2. Multi-objective Optimization Using DE

2.1 The MO Problem

In many practical or real life problems, there are many (possibly conflicting) objectives that need to be optimized simultaneously. Under such circumstances there no longer exists a single optimal solution but rather a whole set of possible solutions of equivalent quality. The field of Multi-objective Optimization (MO) [17 – 19] deals with simultaneous optimization of multiple, possibly competing, objective functions.

2.2 The Differential Evolution (DE) Algorithm

DE [20, 21] is a population-based global optimization algorithm that uses a floating-point (real-coded) representation. It uses crossover (binomial in this case) and mutation operations to optimize a given cost function. For want of space, we avoid mentioning the details of the DE algorithm here and refer the reader to the aforementioned literatures.

2.3 The Multi-objective Variant of DE

We have used the Multi-objective DE (MODE) [4]. MODE was proposed by Xue *et al*. [8]. This algorithm uses a variant of the original DE, in which the best individual is adopted to create the offspring. A Pareto-based approach is introduced to implement the selection of the best individual. If a solution is dominated, a set of non-dominated individuals can be identified and the "best" turns out to be any individual (randomly picked) from this set.

3. Multi-objective Clustering Scheme

3.1 Search-variable Representation and Description of the new algorithm

In the proposed method, for n data points, each d-dimensional, and for a user-specified maximum number of clusters $K_{\max}$, a chromosome is a vector of real numbers of dimension $K_{\max} + K_{\max} \times d$. The first $K_{\max}$ entries are positive floating-point numbers in [0, 1], each of which controls whether the corresponding cluster is to be activated (i.e. to be really used for classifying the data) or not. The remaining entries are reserved for $K_{\max}$ cluster centers, each d-dimensional. For example, the i-th vector is represented as:

$\vec{X}_i(t) =$

$T_{i,1}$		....	$T_{i,K_{\max}}$	$\vec{m}_{i,1}$	$\vec{m}_{i,2}$	.....	$\vec{m}_{i,K_{\max}}$

The *j*-th cluster center in the *i*-th chromosome is active or selected for partitioning the associated dataset if $T_{i,j} = 1$. On the other hand, if $T_{i,j} = 0$, the particular *j*-th cluster is inactive in the *i*-th vector in DE population. Thus the $T_{i,j}$ s behave like control genes.

IF $T_{i,j} = 1$ **THEN** the *j*-th cluster center $\vec{m}_{i,j}$ is **ACTIVE**

ELSE $\vec{m}_{i,j}$ is **INACTIVE.** (1)

Conjunction of GA and DE algorithms:

The Differential Evolution algorithm is applied on the first $K_{\max}$ members of the chromosome (as activated by the corresponding control genes), whereas, the control genes form a binary encoded GA population, which are operated by the Genetic operators of Selection, Crossover and Mutation. Binary tournament selection is employed in this case. The different GA operators are not reiterated here due to space limitations.

Simple generational genetic algorithm pseudo code:

1. Choose initial population
2. Evaluate the fitness of each individual in the population
3. Repeat until termination: (time limit or sufficient fitness achieved)
    1. Select best-ranking individuals to reproduce
    2. Breed new generation through crossover and/or mutation (genetic operations) and give birth to offspring
    3. Evaluate the individual fitnesses of the offspring

Replace worst ranked part of population with offspring.

3.2 Selecting the Objective Functions

Conflict among the objective functions is often beneficial since it guides to globally optimal solutions. In this work we choose the Xie-Beni index XB_q and the FCM objective function J_q as the two objectives. The FCM measure J_q may be defined as:

$$J_q = \sum_{j=1}^{n} \sum_{i=1}^{k} u_{ij}^{q} \cdot d^2(\vec{Z}_j, \vec{m}_i), \; 1 \le q \le \infty \qquad (2)$$

where q is the fuzzy exponent, d indicates a distance measure between the *j*-th pattern vector and *i*-th cluster centroid, and u_{ij} denotes the membership of *j*-th pattern in the *i*-th cluster. The XB index is defined as a function of the ratio of the total variation σ to the minimum separation sep of the clusters. Here σ and sep may be written as:

$$\sigma = \sum_{i=1}^{k}\sum_{p=1}^{n} u_{ip}^2 \cdot d(\vec{m}_i, \vec{Z}_p) \tag{3}$$

and $$sep(Z) = \min_{i \neq j}\left\{d^2(\vec{m}_i, \vec{m}_j)\right\} \tag{4}$$

The XB index is then written as:

$$XB_q = \frac{\sigma}{n \times sep(Z)} = \frac{\sum_{i=1}^{k}\sum_{p=1}^{n} u_{ip}^q \cdot d^2(\vec{m}_i, \vec{Z}_p)}{n \times \min_{i \neq j}\left\{d^2(\vec{Z}_i, \vec{Z}_j)\right\}} \tag{5}$$

Let the set of centers be denoted by $\{\vec{m}_1, \vec{m}_2, ..., \vec{m}_k\}$. The membership value of the *j*-th pattern in *i*-th cluster $u_{ij}, i = 1,2,....k$ and $j = 1,2,....,n$ are computed as:

$$u_{ij} = \frac{1}{\sum_{p=1}^{k}\left(\frac{d(\vec{m}_i, \vec{Z}_j)}{d(\vec{m}_p, \vec{Z}_j)}\right)^{\frac{2}{q-1}}} \tag{6}$$

Note that while computing the u_{ij} s, using equation (12), if $d(\vec{m}_p, \vec{Z}_j)$ is equal to zero for some *p*, then u_{ij} is set to zero for all $i = 1,2,....k$, $i \neq j$, while u_{pj} is set equal to one. Subsequently the centers encoded in a vector are updated using:

$$\vec{m}_p = \frac{\sum_{j=1}^{n}\left(u_{pj}\right)^q \cdot \vec{Z}_j}{\sum_{j=1}^{n}\left(u_{pj}\right)^q} \tag{7}$$

3.3 Avoiding Erroneous Vectors

There is a possibility that in our scheme, during computation of the *XB* or J_q, a division by zero may be encountered. This may occur when one of the selected cluster centers in a DE-vector is outside the boundary of distributions of the data set. To avoid this problem we first check to see if any cluster has fewer than two data points in it. If so, the cluster center positions of this special chromosome are re-initialized by an average computation.

3.4 Selecting the Best Solution from Pareto-front

For choosing the most interesting solutions from the Pareto front, we apply Tibshirani *et al.* Gap statistic [24], a statistical method to determine the number of clusters in a data set.

3.5 Evaluating the Clustering Quality

In this work, the final clustering quality is evaluated using two external measures. Specifically we choose the Adjusted Rand Index [25] (which is a generalization of the Rand index [26]) and the Sihouette index [27]. Silhouette width reflects the

compactness and separation of the clusters. Given a set of data points $Z = \{\vec{Z}_1, ..., \vec{Z}_n\}$ and a given clustering solution $C = \{C_1, C_2, ..., C_k\}$, the silhouette width $s(\vec{Z}_j)$ for each data $\vec{Z}_j$ belonging to cluster C_i indicates a measure of the confidence of belongingness, and it is defined as:

$$s(\vec{Z}_j) = \frac{b(\vec{Z}_j) - a(\vec{Z}_j)}{\max(a(\vec{Z}_j), b(\vec{Z}_j))}. \quad (8)$$

Here $a(\vec{Z}_j)$ denotes the average distance of data point $\vec{Z}_j$ from the other data points of the cluster to which the data point $\vec{Z}_j$ is assigned (i. e. cluster C_i). On the other hand, $b(\vec{Z}_j)$ represents the minimum of the average distances of data point $\vec{Z}_j$ from the data points belonging to clusters C_r, $r = 1,2,..., k$ and $r \neq i$. The value of $s(\vec{Z}_j)$ lies between -1 and +1. Large values of $s(\vec{Z}_j)$ (near to 1) indicate that the data point $\vec{Z}_j$ is well clustered. Overall silhouette index $s(C)$ of a clustering solution $C = \{C_1, C_2, ..., C_k\}$ is defined as the mean silhouette width over all the data points:

$$s(C) = \frac{1}{n} \sum_{j=1}^{n} s(\vec{Z}_j) \quad (9)$$

4 Experimental results

4.1 Datasets used

The experimental results showing the effectiveness of multi-objective DE based clustering has been provided for six artificial and four real life datasets. Table 1 presents the details of the datasets. The real-life datasets are iris, wine, breast-cancer [28] and the yeast sporulation data. The sporulation dataset is available from [31].

4.2 Parameters for the Algorithms

GADE has been used with 40 parameter vectors in each generation and each run of each algorithm was continued for 100 generations. The value of scale factor F is a random value between 0.5 and 1. The other parameters for the multi-objective GA (NSGA II) based clustering are fixed as follows: number of generations = 100, population size = 50, crossover probability = 0.8, mutation probability $= \frac{1}{Chromosome\_length}$. Please note that GADE and the NSGA II use the same parameter representation scheme. Clustering with MOCK was performed with the source codes available from [32].

4.3 Presentation of Results

The mean Silhouette index values of the best-of-run solutions provided by six contestant algorithms over the 10 datasets have been provided in Table 2. The best entries have been marked in boldface in each row. Table 3 enlists the adjusted rand index values except for Yeast sporulation data as no standard nominal classification is known for this dataset.

4.4.4 Significance and Validation of Microarray Data Clustering Results

In this section the best clustering solution provided by different algorithms on the sporulation data of yeast has been visualized using the cluster profile plot (in parallel coordinates[30]) in MATLAB 7.0.4 version. It is a common way of visualizing high-dimensional geometry. Cluster profile plots (in parallel coordinates) of seven clusters for the best clustering result (provided by GADE) on yeast sporulation data has been shown in Figure 1. The blue polylines indicate the member genes within a cluster while the black polyline indicates the centroid of that gene. The heatmap and fatigo results may be obtained from [33].

Table 1. Details of the datasets used.

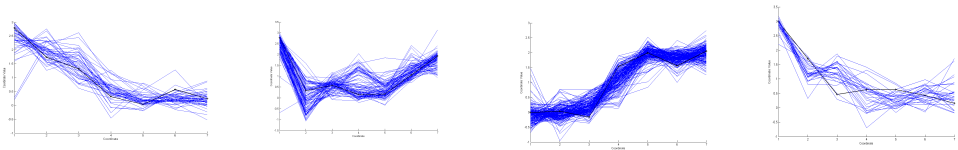
Dataset	**Number of points**	**Number of clusters**	**Number of Characteristics**
Dataset_1	900	9	2
Dataset _2	76	3	2
Dataset _3	400	4	3
Dataset _4	300	6	2
Dataset _5	500	10	2
Dataset_ 6	810	3	2
Iris	150	3	4
Wine	178	3	13
Breast-Cancer	683	2	9
Yeast Sporulation	474	7	7

Table 2. Mean value of sil index found and standard deviations (in parentheses) by contestant algorithm over 30 independent runs on ten datasets.

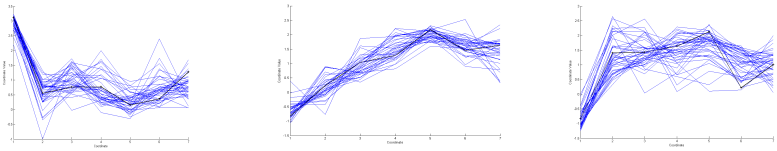
Dataset	**Algorithms Compared**					
	GADE		**NSGA II**		**MOCK**	
	k	Silhouette Index	*k*	Silhouette Index	*k*	Silhouette Index
Dataset_1	**9.12 (1.46)**	**0.735312 (0.254134)**	9.37 (1.72)	0.669317 (0.0892)	8.52 (2.81)	0.66342 (0.0736)
Dataset_2	**3.36 (0.65)**	**0.664993 (0.123610)**	3.16 (0.072)	0.654393 (0.00927)	3.33 (1.03)	0.658921 (0.004731)
Dataset_3	4.14 (0.36)	**0.872521 (0.127479)**	3.57 (0.51)	0.765691 (0.005686)	3.78 (1.25)	0.768419 (0.006721)
Dataset_4	6.04 (0.25)	0.705079 (0.115616)	6.28 (0.46)	0.827618 (0.02871)	**6.08 (0.51)**	**0.832527 (0.007825)**
Dataset_5	9.24 (3.89)	**0.771040 (0.042776)**	12.43 (0.939)	0.768379 (0.005384)	10.41 (0.80)	0.769342 (0.006208)
Dataset_6	5.19 (0.93)	**0.792000 (0.208000)**	4.65 (1.58)	0.642091 (0.002833)	5.16 (0.38)	0.640957 (0.008349)
Iris	2.31 (0.76)	0.429655 (0.331443)	2.16 (1.06)	0.566613 (0.082651)	**3.05 (0.37)**	**0.6003725 (0.005129)**
Wine	**3.16 (0.46)**	**0.582197 (0.00427)**	3.88 (0.67)	0.5767342 (0.009415)	3.59 (0.46)	0.576834 (0.000812)
Breast Cancer	**2.08 (0.38)**	**0.648297 (0.00734)**	2.57 (0.60)	0.6004642 (0.004561)	2.10 (0.53)	0.626719 (0.01094)
Yeast Sporulation	**7.08 (0.12)**	**0.641630 (0.212575)**	7.22 (0.68)	0.641306 (0.04813)	6.67 (0.857)	0.613567 (0.005738)

Table 3. Mean value of adjusted rand index found and standard deviations (in parentheses) by contestant algorithm over 30 independent runs on ten datasets.

Dataset	Algorithms Compared		
	GADE	**NSGA2**	**MOCK**
Dataset_1	**0.884288 (0.101020)**	0.802180(0.004782)	0.810934 (0.0059348)
Dataset_2	**0.951535 (0.179265)**	0.9378123(0.006821)	**0.946547 (0.004536)**
Dataset_3	0.850030 (0.152226)	**0.963841(0.0046719)**	0.878732 (0.0712523)
Dataset_4	0.785995(0.137284)	0.957818 (0.004678)	**0.978761** (0.006734)
Dataset_5	0.788450 (0.019142)	**0.947641 (0.006646)**	0.9454568 (0.0012043)
Dataset_6	0.692516 (0.168323)	0.881395 (0.056483)	**0.910294 (0.016743)**
Iris	**0.843862 (0.076887)**	0.715898 (0.005739)	0.786574 (0.075763)
Wine	**0.875849 (0.0087642)**	0.828645(0.0074653)	0.864764 (0.0034398)
Breast Cancer	**0.956456 (0.0053)**	0.944236(0.006521)	0.9465731 (0.006748)



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4



(e) Cluster 5 (f) Cluster 6 (g) Cluster 7

5. Conclusions

This paper compared and contrasted the performance of GADE in an automatic clustering framework with two other prominent multi-objective clustering algorithms. The multi-objective GADE-variant used the same variable representation scheme. Tables 2 to 4 indicate that GADE was usually able to produce better final clustering results as compared to MOCK or NSGA II in terms of both adjusted Rand index and Silhouette index when all the algorithms were let run for an equal number of generations. Future research may extend the multi-objective GADE-based clustering schemes to handle discrete chromosome representation schemes that no longer depend on cluster centroids and thus are not biased in any sense towards spherical clusters.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review", *ACM Computing Surveys*, vol. 31, no.3, (1999) 264—323.

[2] R. Xu and D. Wunsch, *Clustering*, Series on Computational Intelligence, IEEE Press, 2008.

[3] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery", *IEEE Transactions Geoscience and Remote Sensing,* 2006.

[4] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering", *IEEE Transactions on Evolutionary Computation*, 11(1):56-76, 2007.

[5] H. A. Abbass and R. Sarker, "The pareto differential evolution algorithm", *International Journal on Artificial Intelligence Tools*, 11(4):531-552, 2002.
[6] F. Xue, A. C. Sanderson, and R. J. Graves, "Pareto-based multi-objective differential evolution", in *Proceedings of the 2003 Congress on Evolutionary Computation (CEC'2003)*, volume 2, pages 862–869, Canberra, Australia, 2003, IEEE Press.
[7] T. Robic and B. Filipic, "DEMO: Differential Evolution for Multiobjective Optimization", In C. A. Coello Coello, A. H. Aguirre, and E. Zitzler, editors, *Evolutionary Multi-Criterion Optimization, Third International Conference, EMO 2005*, pages 520–533, Guanajuato, Mexico, 2005, Springer LNCS Vol. 3410, 2005.
[8] A. W. Iorio and X. Li, "Solving rotated multi-objective optimization problems using differential evolution", in *AI 2004: Advances in Artificial Intelligence, Proceedings*, pages 861–872, Springer-Verlag, LNAI Vol. 3339, 2004.
[9] S. Paterlinia, T. Krink, "Differential evolution and particle swarm optimisation in partitional clustering", *Computational Statistics & Data Analysis*, Volume 50, Issue 5, 1220-1247, 2006.
[10] M. Omran, A. P. Engelbrecht and A. Salman, "Differential evolution methods for unsupervised image classification", Proceedings of *Seventh Congress on Evolutionary Computation* (CEC-2005). IEEE Press, 2005.
[11] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm", *IEEE Transactions on Systems Man and Cybernetics - Part A*, Vol. 38, No. 1, pp. 1-20, January 2008.
[12] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, 2002.
[13] D.W. Corne, J.D. Knowles, and M.J. Oates, "The pareto-envelope based selection algorithm for multiobjective optimisation", in M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H-P. Schwefel, (eds.) *Parallel Problem Solving from Nature - PPSN VI*, Springer Lecture Notes in Computer Science, pp. 869—878, 2000.
[14] X. Xie and G. Beni, "Validity measure for fuzzy clustering", *IEEE Trans. Pattern Anal. Machine Learning*, Vol. 3, pp. 841–846, (1991).
[15] J. C. Bezdek, "Cluster validity with fuzzy sets", *Journal of Cybernetics*, (3) 58–72, (1974).
[16] S. Chu *et al*. "The transcriptional program of sporulation in budding yeast", *Science*, 282, 699–705, 1998.
[17] Y. Sawaragi, H. Nakayama, and T. Tanino, "Theory of multiobjective optimization" (vol. 176 of *Mathematics in Science and Engineering*). Orlando, FL: Academic Press Inc., 1985.
[18] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, 2001.
[19] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, Springer, 2007.
[20] R Storn and K Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces", *Journal of Global Optimization*, 11(4) (1997) 341–359.
[21] R. Storn, K. V. Price, and J. Lampinen, *Differential Evolution - A Practical Approach to Global Optimization,* Springer, Berlin, 2005.
[22] C. A. Mattson, A. A. Mullur, and A. Messac, "Smart Pareto filter: Obtaining a minimal representation of multiobjective design space," *Eng. Optim.*, vol. 36, no. 6, pp. 721–740, 2004.
[23] J. Branke, K. Deb, H. Dierolf, and M. Osswald, "Finding knees in multi-objective optimization," in *Proc. 8th Int. Conf. Parallel Problem Solving From Nature*, pp. 722–731, 2004.
[24] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the Gap statistic," *J. Royal Statist. Soc.: SeriesB (Statistical Methodology),* vol. 63, no. 2, pp. 411–423, 2001.
[25] L. Hubert and P. Arabie, "Comparing partitions", *Journal of Classification*, 193–218, 1985.
[26] W. M. Rand, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*, 66, 846–850, 1971.
[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[28] C. Blake, E.Keough and C.J.Merz, UCI repository of machine learning database (1998). http://www.ics.uci.edu/~mlearn/MLrepository.html
[29] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Second Edition, Elsevier Academic Press, 2003.
[30] D. A. Keim and H.-P. Kriegel, "Visualization techniques for mining large databases: a comparison", *IEEE Transactions on Knowledge and Data Engineering*, v.8 n.6, p.923-938, December 1996.
[31] http://cmgm.stanford.edu/pbrown/sporulation
[32] http://dbkgroup.org/handl/mock/
[33] http://swagatamdas19.googlepages.com/