Pertanika J. Sci. & Technol. 28 (1): 141 - 162 (2020)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Clustering with Modified Mutation Strategy in Differential Evolution

Seema Patil[1]* and Anandhi Rajamani Jayadharmarajan[2]

[1]Department of Computer Science and Engineering, The Oxford College of Engineering, Bangalore- 560068 and affiliated to Visvesvaraya Technological University, Belgavi, Karnataka, India
[2]Department of Information Science & Engineering, New Horizon College of Engineering, Bangalore-560103 and affiliated to Visvesvaraya Technological University, Belgavi, Karnataka, India

ABSTRACT

This paper proposes a clustering approach based on Modified Mutation strategy in the Differential Evolution (MMDE). Differential evolution is an evolutionary computation technique used for optimization. Though DE is very efficient, it sometimes suffers from the issue of slow convergence and the difficulty of achieving a global solution. To overcome these issues, in this paper, a modified mutation method was developed, which maintained the balance between exploration and exploitation. The objectives of modification were to achieve a higher rate of convergence and to obtain better cluster efficiency. The proposed form of modification had been applied on probabilistic environment to define the differential vector through randomly selected members and to obtain the best solution. Over the number of benchmark dataset, clustering efficiency had been estimated and compared with Conventional Differential Evolution (CDE) as well as Particle Swarm Optimization. The proposed method had been tested on a number of benchmark datasets. Experimental results had shown that MMDE had better and consistent clustering efficiency when compared to Conventional Differential Evolution (CDE) and Dynamic Weighted Particle Swarm Optimization (DWPSO).

Keywords: Clustering, convergence, differential evolution, mutation, particle swarm optimization

ARTICLE INFO

Article history:
Received: 11 July 2019
Accepted: 27 November 2019
Published: 13 January 2020

E-mail addresses:
theseema@gmail.com (Seema Patil)
rjanandhi@hotmail.com (Anandhi Rajamani Jayadharmarajan)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Data based knowledge offer numerous opportunities in various practical applications like bioinformatics, engineering, biology, healthcare, medicine, prediction analysis, crime forecasting and computing techniques. The tremendous growth of data-based knowledge in scientific studies has presented a lot of challenges before the researchers to extract useful information using traditional database techniques. Hence effective mining methods are essential to discover the implicit knowledge from huge database.

This knowledge extraction is done with the help of data mining techniques such as classification and clustering. Clustering is an important task of combining various population or data points into clusters. Clustering performs grouping of similar points. It is iterative process to discover the knowledge which involves major trial and failure. The clustering process does not require any kind of feedback to perform similarity of data points, it is self-organized. Clustering using PSO defines a new Swarm Intelligence (SI) for partitioning any datasets into an optimal number of groups through a single run of optimization. SI is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from biological examples such as swarming, flocking and herding behavior in vertebrates.

Data clustering is a popular approach of automatically finding classes, concepts, or groups of patterns. Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, and swarms of bees, and even in human social behavior. Data clustering using PSO can be used to find the centroids of user specified number of clusters. DE is one of the most powerful algorithms available in the community and it is being used for various practical purposes. DE works on the steps of defining differential vector, mutation vector, crossover, selection and finally with termination step, any further enhancement in the fundamental structure of DE will help to improve the quality of performance.

This work proposes the method for clustering based on differential evolution. Even though DE is very efficient, sometimes it suffers from the issue of slow convergence and the difficulties in achieving a global solution. To overcome these, balance between exploration and exploitation has been maintained by adding the two modules in the conventional DE. To increase the level of exploitation, under the probabilistic mode, selection between best and randomly selected member takes place. The Differential vector made by best solution, delivers fast change in the solution and results in a faster convergence. The multi-culture approach helps in exploration of new and efficient solutions. Gathering and selection of solutions from different environments will maintain the diversity in the population.

Related Work

Das et al. (2008) used Differential Evolution for automatic clustering of large unlabeled data sets. Gupta and Saini (2018) proposed a new efficient clustering approach which was applied on k harmonic means (KHM) by using PSO. The local optimum problem of KHM was overcome by PSO. Also, fuzzy logic was used to control the various parameters of PSO. Nerurkar et al. (2018) had achieved the global optima on clustering by making use of two validation indices criteria. These indices were simple and robust against other outliers and showed best clustering that had lower computation cost and parallel execution and faster convergence.

Wang et al. (2018) combined PSO and DE approach by taking velocity update of PSO and mutation parameter of DE to generate the new population. The DE re-mutation, crossover and selection were performed throughout the optimization process to get good results. This approach gave the best result when compared with inertia weight PSO and comprehensive learning PSO and basic DE. Zhu et al. (2018) discussed complications associated with K-means clustering algorithm and proposed the concept of centroid all rank distance. Liu et al. (2018) presented an efficient and intelligent DDC algorithm that helped to overcome the difficulties associated with density and delta distance clustering (DDC) when data derived from the two indicators were large.

Yi et al. (2018) had presented a robust recommendation algorithm based on kernel principal component analysis and fuzzy c-means clustering. Kuo and Zulvia (2019) presented a variation of differential evolution (DE) algorithm to solve an automatic clustering problem. Tran et al. (2015) described the new improved approach of PSO by improving the diversity mechanism and mutation operator to employ new neighborhood search strategy. These new approaches were tested on well-defined benchmark data sets. Jiau et al. (2006) presented a hierarchical clustering algorithm based on matrix partitioning.

MATERIALS AND METHODS

Modified Mutated DE (MMDE)

DE is one of the most powerful algorithms in which the formation of Differential vector is the central part that defines the quality of final solution. The existing method of DE based on random member selection slows down the convergence and results in a suboptimal solution.

To increase the convergence speed of DE, a new approach in mutation operation has been presented. It has two possibilities of differential change under the probabilistic environment. In the first case, differential change is defined through best member and random selected member. While in second case, three random members are selected to define the differential change. A threshold value is defined to determine the selection of differential change type. Best member based differential change generates faster change, while the random member-based selection prevents from suboptimal convergence. The pseudo code for applied mutation strategy is shown below:

```
1.      Define the parameter value for:
   Popsz← population size,     Mf   ← mutation rate,     Cf    ←cross-over rate,
   K     ← No. of Clusters,      Thr ←Define a Threshold value
  Dm← problem dimension (k *No. of data attributes)

2. Initialize the population:
   For i=1:Popsz
        POP (i, :) ← [Select 'K' Random sample of data from data set and convert into an array];
   End
3. While termination doesn't occur, {do
         For i=1: Popsz
             X ← POP(i,:);
             r ←U [0, 1]; a random number generated through uniform distribution in range of [0, 1];

              if  r < Thr
                • Select two members' m1 & m2 randomly from population
                •Select best member BM from population
                •Mutation vector defined as: Mv = m1+ Mf* [ BM- m2];
               Else
                •Select three members m1, m2 & m3 randomly from population
                • Mutation vector defined as: Mv= m1+ Mf*[m2-m3]
             End

              rv ←generate a random vector of size [1 l] by U [0, 1];

              For j=1: dm

                   if  rv(j) < Cr
                        Ox(j)  ← Mv(j);
                   Else
                        Ox(j)  ← X(j);
                   End

              End

              Select the better one among parent and offspring
                 If  f_fitness(Ox) > f_fitness (X)
                     NPOP(i,:)←  Ox;
                 Else
                      NPOP(i,:)←  X;
                  End;
    End
Bs←Best solution from NPOP
If (termination doesn't occur)
          Go to step 3
Else
    Final solution← Bs;
End
```

In this proposed work *Thr* is considered as 0.2. Threshold value should not be high otherwise population will lose the diversity soon.

Multi-domain-based DE

A multi-culture concept called "Multi-culture modified mutation Differential Evolution" has been developed to evolve the individual population independently and later to exploit the population to form a better community, which allows an efficient search of the solution space. This approach finds its inspiration from the present human society, where two things can happen at the fundamental level (i) the independent existence of a number of separate populations, each progressing under the same environment up to a certain period of time, (ii) a number of selected individuals belonging to different population, forms a new population to achieve desired objectives.

Rather than working under a monoculture formed by one population as in the conventional PSO, in this paper a multicultural environment has been considered, where a number of different environments are independently created by a different set of population. This study considered population samples that had undergone independent social evolution and among all from the diverse population samples the best individual were selected to finish the task. This was a dual stage process where first stage found some potential solution discovered from different regions of solution space, and later in the second phase, each individual contributed more efficiently to find a global solution. Even with the small size of the population, the proposed method had achieved better quality solution with a very high degree of consistency.

In the working principle of MMDE, population (POP) represents the initial random population, that evolves through the DE process, individually and independently, undergoing fewer iterations, and creates the multi-culture new population (NPOP). Even though the process of creating the NPOP is same for all POP, because of difference in leadership and different community environment, each NPOP has different characteristics. Through the fitness-based selection process, among all members of NPOP, better members were selected to form a new population (SPOP), which had the same size as initial POP. In SPOP, there are a number of good candidates, which are different and have higher fitness value, hence high level of diversity exists in the SPOP. Finally, to obtain the Final Population (FPOP), MMDE applied over SPOP, till the terminating criteria are met.

Datasets Used for Experiment

Experiment was conducted on the three benchmark data sets of UCI repository (Dua & Graff, 2019). Wine, Iris, and Glass datasets had been considered to analyze and compare the performance of proposed method with the evolutionary methods. Table 1 shows the details of Wine, Iris, and Glass datasets.

Table 1
Description of datasets

Data set	No of attributes	No of instances	Type of data
Wine	13	178	Multivariate
Iris	4	150	Multivariate
Glass	10	214	Multivariate

RESULTS AND DISCUSSION

This section discusses the results obtained from various evolutionary methods. Parameters under discussion are size of population, mutation, crossover rate, and number of iterations. Lower population size will have lesser diversity and low exploration rate, which may result in convergence with suboptimal solution. Higher population size may lead to very slow convergence. So practically, depending upon the kind of application, a population size of 50 to 150 is generally considered.

In practical applications, it was observed that low value of mutation and crossover rate caused slower convergence, while high mutation rate led to travelling the same area of search landscape. It also observed that high crossover rate caused loss of diversity and led to less exploration. It was also observed that very high value of crossover would cause loss of diversity and exploration. Therefore, a moderate value in the range of 0.3 to 0.7 was preferred for both mutation and crossover rate and the mutation rate value must be lesser than the crossover rate value.

In the algorithm design process, instead of opting for self-termination it is better to give enough opportunity to the algorithm to come up with optimum results; this helps in evaluating the strength of the algorithm. During the process, more diversion characteristics had been observed from 100 to 200 iterations, while stability was observed from 300 to 400 iterations and which was maintained up to 600 iterations, which could be clearly observed in the figures shown below. Hence, 600 iterations were good enough to define the stability of algorithm performance.

Therefore, in this experiment, the size of population had been considered as 100, mutation rate and crossover rate as 0.4 and 0.5 respectively and the allowed number of iterations as 600.

The performances of all three approaches (DWPSO, CDE, MMDE) have been represented in a tabular format in terms of no of trials, correctly placed data samples in the clusters, number of data samples placed wrongly, cluster efficiency and total intra cluster distance value.

In the first part, only the MMDE had been applied and performances had been obtained for 5 independent trials for ‘Wine’, ‘Iris’ and ‘Glass’ datasets. Comparison had been made with Conventional DE (CDE) and Dynamic Weighted PSO (DWPSO).

In second part, multidomain based experiment had been included with MMDE and performances had been estimated over "Glass" data set. Experimental process had been developed in the MATLAB version 7.1 environment.

Dataset: Wine Data

There are total 178 set of data carrying 3 clusters. Each data contains 13 attributes.

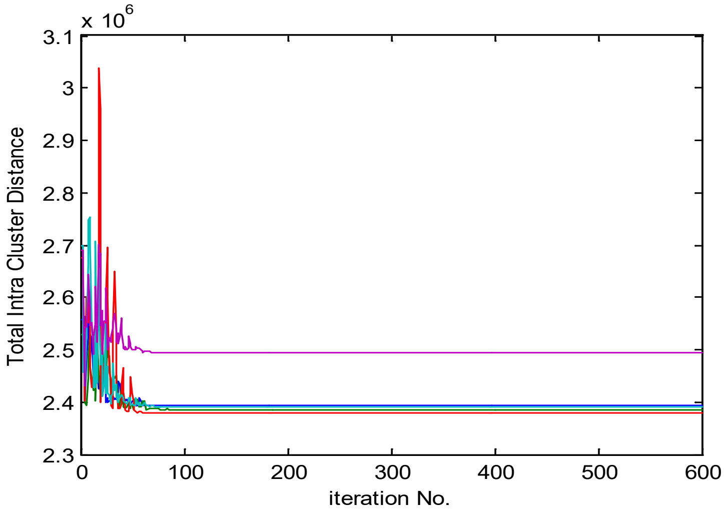

Figure 1. DWPSO based convergence in 5 trials for wine data set

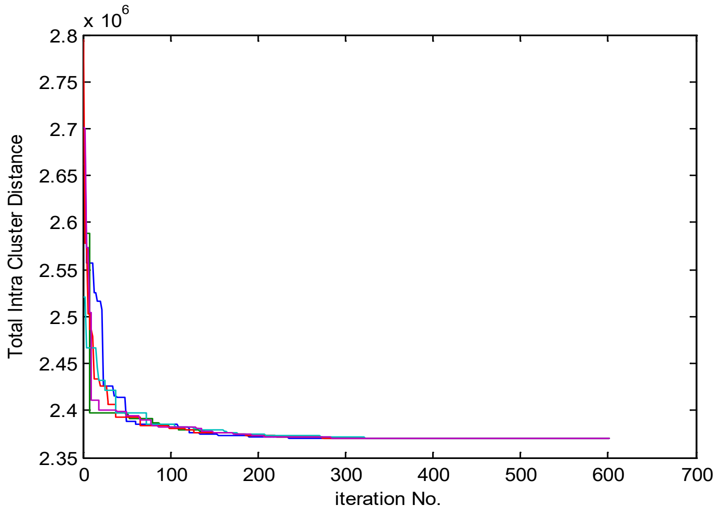

Figure 2. CDE based convergence in 5 trials for wine data set

Table 2

Mean Performance over 5 trials by different algorithm over wine data set

	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value 1.0e+006 *
DWPSO	125	53	70.22	2.4088e+006
CDV	125	53	70.22	2.3707e+006
MMDV	125	53	70.22	2.3707e+006

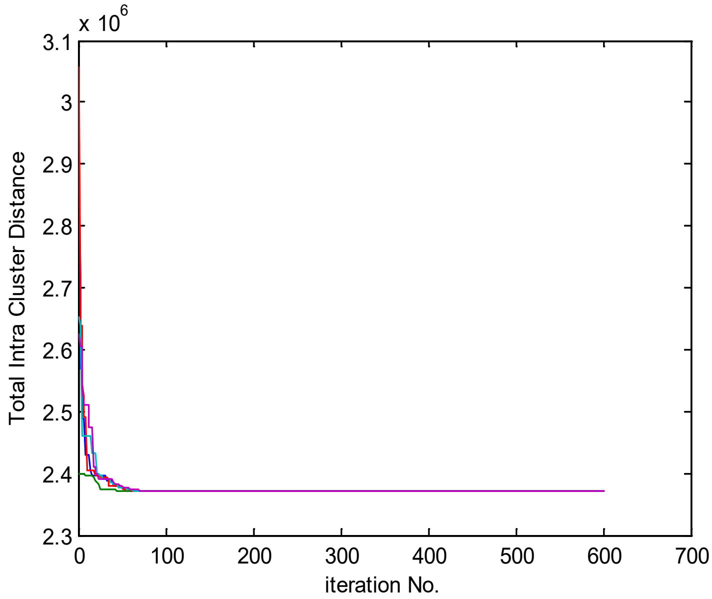

Figure 3. MMDE based convergence in 5 trials for wine data set

Table 3

Centroid position for wine data

	Centroids of Wine data set								
C1	3.0351	3.0067	3.0065	3.0541	3.2816	3.0057	3.0043	3.0108	3.0041
C2	3.0375	3.0051	3.0065	3.0462	3.2867	3.0078	3.0081	3.0008	3.0051
C3	3.0339	3.0067	3.0062	3.0565	3.2508	3.0057	3.0048	3.0010	3.0040

	Centroids of Wine data set			
C1	3.0154	3.0024	3.0067	4.9797
C2	3.0154	3.0029	3.0084	6.2486
C3	3.0111	3.0024	3.0067	4.2455

The performances obtained under 5 independent trials by different algorithms are shown in Table 2. It can be observed that all the three algorithms have nearly the same performances; while the distance value is marginally greater for DWPSO. The obtained centroid values by MMDE for the 1st trial are shown in Table 3.

The convergence characteristics for DWPSO over wine data in 5 independent trials have been shown in Figure 1. It can be observed that, good amount of diversity existed in their convergence characteristics and mean convergence value obtained was around 2.4088e+006, which is shown in Table 2. Over the same data set, CDE had been applied for 5 independent trials and obtained convergence characteristics are shown in Figure 2.

It can be observed that, nearly same convergence path had appeared over different trials and the obtained final convergence value for total intra cluster distance was around 2.3707e+006 which was substantially less compared to the value obtained by DWPSO. To get clarity on advantages of proposed MMDE, the experiment has been repeated over wine data with MMDE and obtained convergence characteristics have been shown in Figure 3.

It can be observed that faster convergence, with excellent reliability feature, has been achieved compared to both DWPSO and CDE. To demonstrate the relative comparison between DWPSO, CDE and MMDE, their mean performances over 5 trials have been plotted on a graph as shown in Figure 4. It is observed from Table 2 that, the mean intra cluster distance for DWPSO and MMDE were nearly same. However, convergence of MMDE occurred around 100th iteration, while DWPSO took more than 200 iterations to converge.

Dataset: IRIS Data

Iris dataset contains a total of 150 data sets and each data has 4 attributes. Three different global clusters exist in the dataset. The performances over Iris data by different algorithms for a number of trials have been analyzed. Figure 5 depicts the performance of DWPSO in 5 trials for Iris dataset.

It can be observed that, there was a consistency in performance in all the trials. While initially more uncertainty to explore the optimal solution has been observed, later after around the 80th iteration, optimal solution has been explored smoothly. The performance by DWPSO and MMDE on Iris dataset are shown in Figure 6 and in Figure 7 respectively.

Diverse convergence has been observed in the beginning, later smooth convergence has been observed. CDV is the vector used by CDE and MMDV is the vector used by MMDE. The mean convergence characteristics comparison of DWPSO, CDE and MMDE has been shown in Figure 8. Intra cluster distance obtained for DWPSO, CDE and MMDE has been shown in Table 4, 5 and 6 respectively. The obtained centroid values have been shown in Table 7. It is observed that mean distance obtained by MMDE is less compared to distance obtained by CDE and DWPSO.

Table 4
DWPSO performance over Iris data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	134	16	89.33	79.3157
2	134	16	89.33	80.2949
3	133	17	88.67	79.4755
4	136	14	90.67	83.2333
5	133	17	88.67	79.7068
Mean	**134**	**16**	**89.33**	**80.4052**

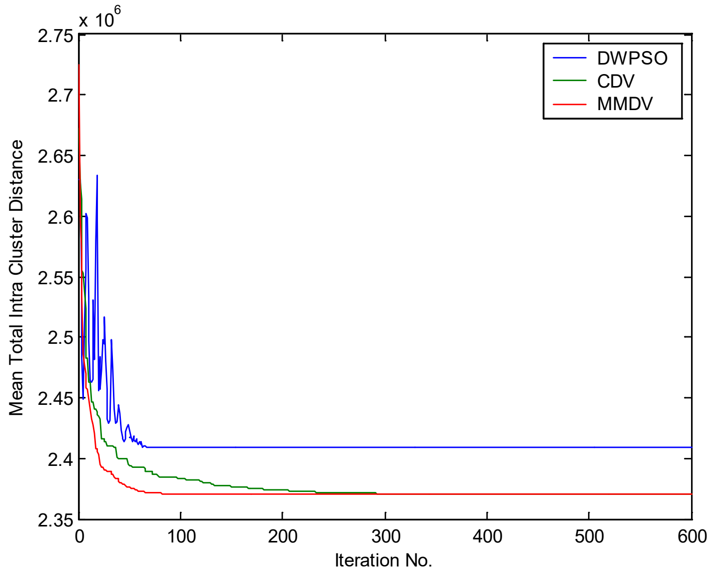

Figure 4. Mean convergence comparison for wine data set

Table 5
CDE performance over Iris data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	134	16	89.33	79.2028
2	134	16	89.33	78.9563
3	133	17	88.67	79.1462
4	134	16	89.33	79.2389
5	134	16	89.33	78.9430
Mean	**133.8**	**16.2**	**89.2**	**79.0974**

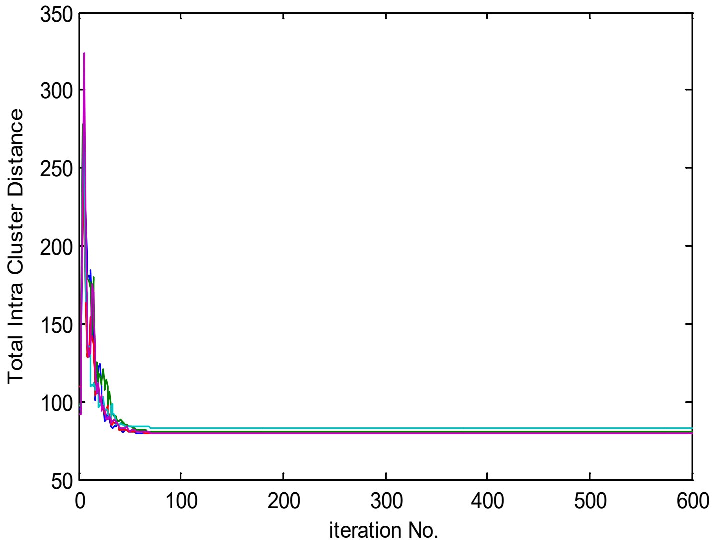

Figure 5. DWPSO based convergence in 5 trials for Iris data set

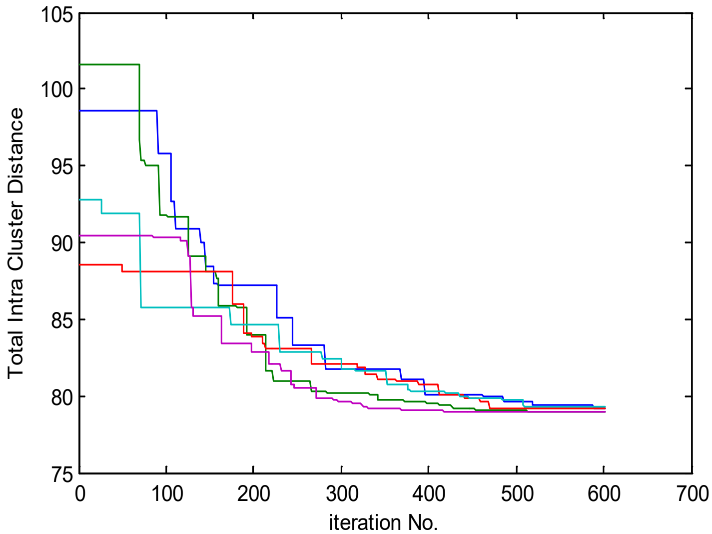

Figure 6. CDE based convergence in 5 trials for Iris data set

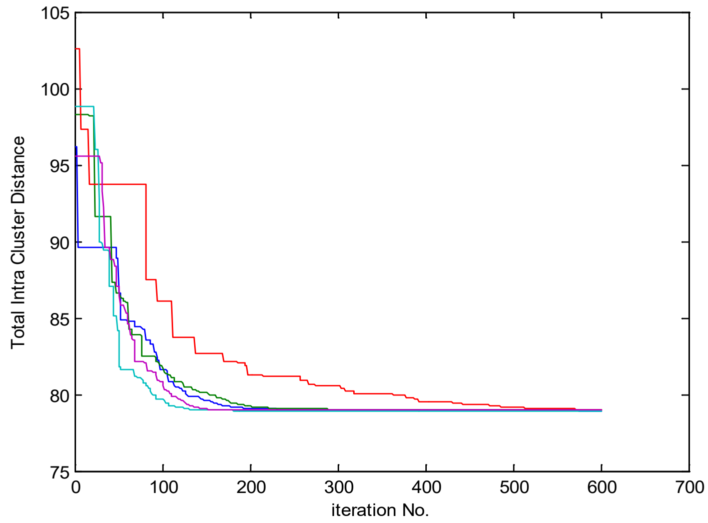

Figure 7. MMDE based convergence in 5 trials for Iris data set

Table 6

MMDE performance over Iris data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	134	16	89.33	78.9471
2	134	16	89.33	78.9631
3	134	16	89.33	79.0133
4	134	16	89.33	78.9454
5	134	16	89.33	78.9494
Mean	**134**	**16**	**89.33**	**78.9637**

Table 7

Centroids value for Iris data set

Centroids of IRIS Dataset				
C1	5.8863	2.7456	4.3731	1.4115
C2	5.0173	3.4385	1.4452	0.2704
C3	6.8326	3.1128	5.7640	2.0469

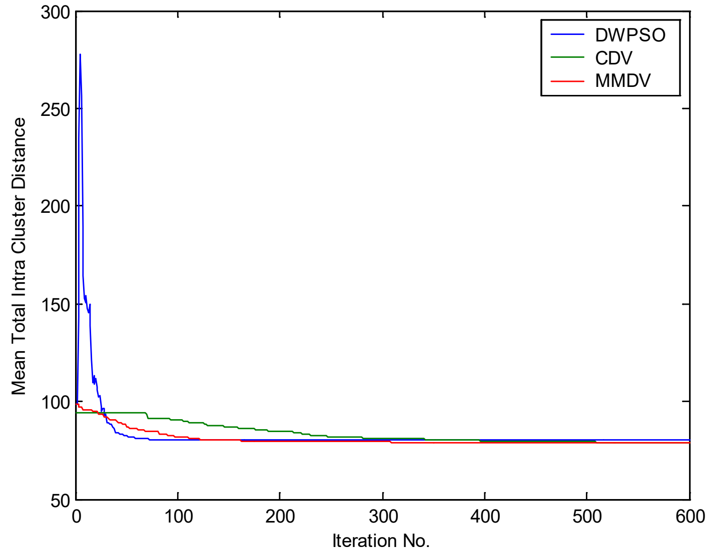

Figure 8. Mean convergence comparison for Iris data set

Dataset: Glass Data

This data set contains total 214 data set. Each data set carried 10 attributes and 6 clusters exists.

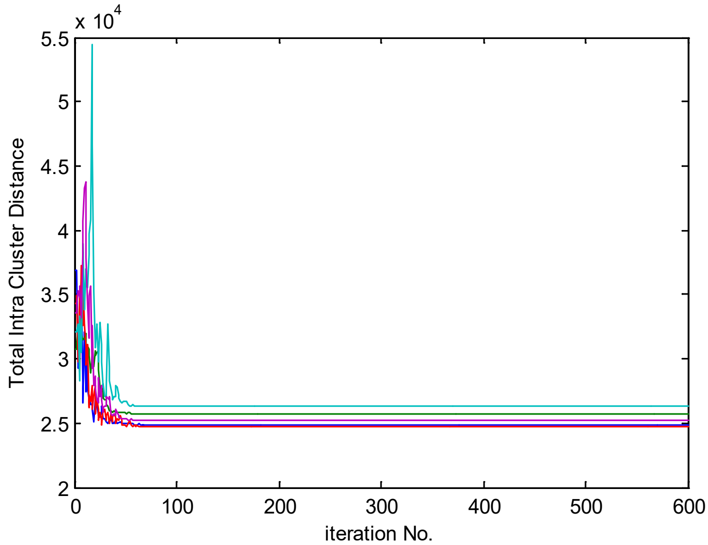

Figure 9. DWPSO based convergence in 5 trials for Glass data set

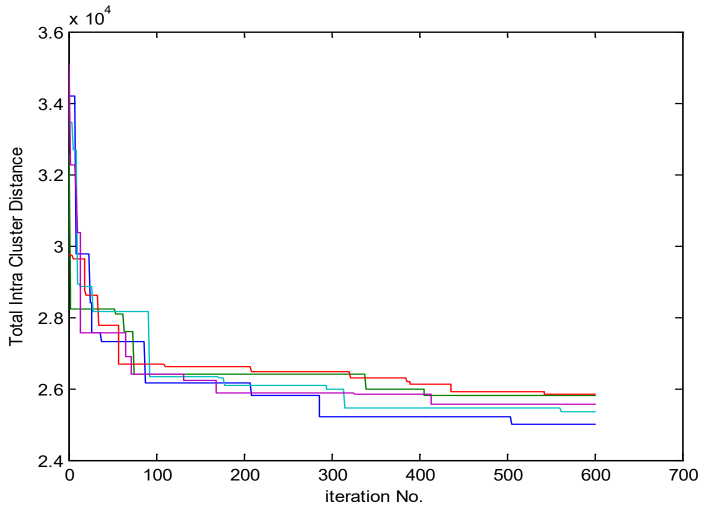

Figure 10. CDE based convergence in 5 trials for Glass data set

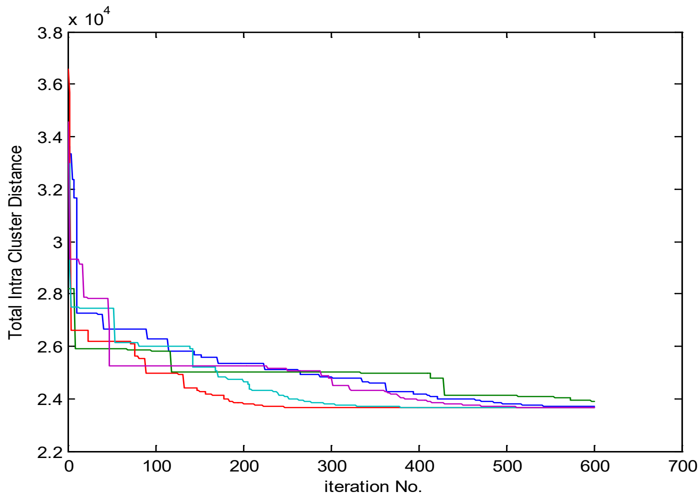

Figure 11. MMDE based convergence in 5 trials for Glass dataset

Table 8

DWPSO performance over Glass data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	183	31	85.51	2.4897 e+004
2	189	25	88.32	2.5737 e+004
3	178	36	83.18	2.4721 e+004

Table 8 *(Continued)*

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
4	184	30	85.98	2.6271 e+004
5	188	26	87.85	2.5209 e+004
Mean	**184.4**	**29.6**	**86.17**	**2.5367e+004**

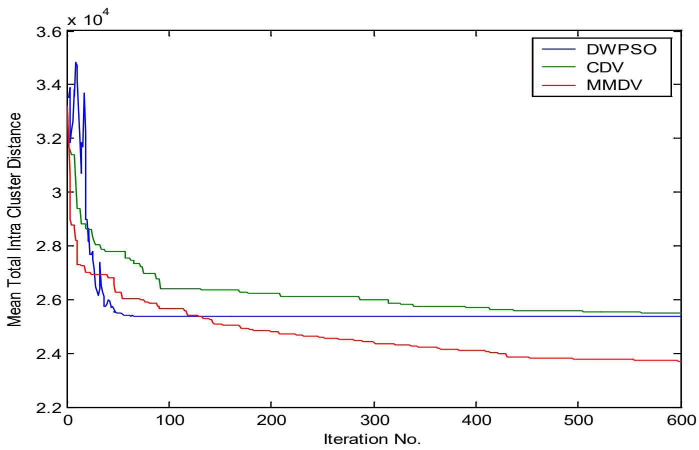

Figure 12. Mean convergence comparison for Glass data set

Table 9

CDE performance over Glass data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	187	27	87.38	2.4990 e+004
2	187	27	87.38	2.5797 e+004
3	187	27	87.38	2.5850 e+004
4	189	25	88.32	2.5368 e+004
5	184	30	85.98	2.5546 e+004
Mean	**186.8**	**27.2**	**87.29**	**2.5510e+004**

Table 10

MMDE performance over Glass data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	183	31	85.51	2.2950 e+004
2	189	25	88.32	2.3174 e+004
3	178	36	83.18	2.3401e+004
4	184	30	85.98	2.4461 e+004
5	188	26	87.85	2.3604 e+004
Mean	**184.4000**	**29.6000**	**86.17**	**2.3518e+004**

Table 11

Centroids value for Glass data set

	Centroids of Glass data set						
C1	166.0782	2.4471	13.7061	3.5266	2.2563	73.3031	2.4611
C2	198.4844	2.5638	16.2827	3.2212	2.7751	73.5565	1.7972
C3	54.2369	2.1344	14.2542	4.4666	1.9043	72.6730	1.0457
C4	18.5031	2.1863	13.2582	4.4278	1.5191	74.4194	1.3567
C5	129.9205	0.8875	13.9521	4.3390	2.7228	75.5818	0.9168
C6	91.0957	2.8459	14.1901	3.6017	2.9122	72.2789	0.9257

	Centroids of Glass data set		
C1	10.7421	-0.1976	0.5747
C2	9.9803	1.6024	-0.1853
C3	9.7003	1.4352	0.2335
C4	10.2181	0.4565	10.1096
C5	8.7067	1.4468	1.4522
C6	10.0617	0.7071	1.1787

Convergence characteristics obtained for DWPSO, CDE and MMDE on glass dataset over 5 independent trials have been shown in Figures 9, 10 and 11 respectively. It is observed that DWPSO convergences faster than CDE. For DWPSO variation is observed in all the trials at the beginning, after 500 iterations, smooth convergence is obtained. Comparative mean convergence for DWPSO, CDE and MMDE is shown in Figure 12. Intra cluster distance obtained for DWPSO, CDE and MMDE is shown in Table 8, 9 and 10 respectively. The obtained best centroid value has also been shown in Table 11. It is observed that mean distance obtained by MMDE is less compared to distance obtained by CDE and DWPSO. MMDE has shown improved tendency of convergence with iteration when compared to DWPSO and CDE.

Multidomain Based MMDE

For the first stage, 10 independent population were considered to maintain the diversity. It can be observed that within 50 iterations each population has evolved in different manner. With this diversity, intra cluster distance of 720 to 775 was obtained. All the evolved population was combined to form the 2nd stage population. The diversity introduced has resulted in achieving minimal value of intra cluster distance.

For the 5 independent trials, the obtained convergence in 1st stage and 2nd stage is shown in Figure 13 and Figure 14 respectively. It can be observed that multidomain MMDE has shown remarkable improvement and mean intra cluster distance value of 702.0192 MMDE has been obtained. Also, better cluster efficiency of 87.48% has been obtained as displayed in Table 12. The corresponding centroid values obtained by multidomain MMDE has been presented in Table 13.

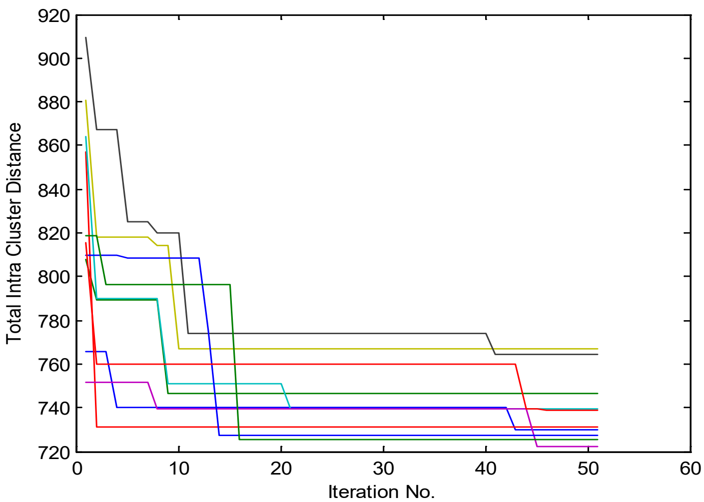

Figure 13. Convergence characteristics in 1st Stage for multidomain MMDE

Table 12

Multidomain MMDE performance over Glass data

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	188	26	87.85	695.5811
2	188	26	87.85	694.0454

Table 12 (*Continued)*

Trial No.	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
3	189	25	88.32	707.4350
4	190	24	88.79	697.8723
5	181	33	84.58	715.1624
Mean (Std.Dev)	**187.2 (3.5637)**	**26.8 (3.5637)**	**87.48 (0.1252)**	**702.0192 (9.042)**

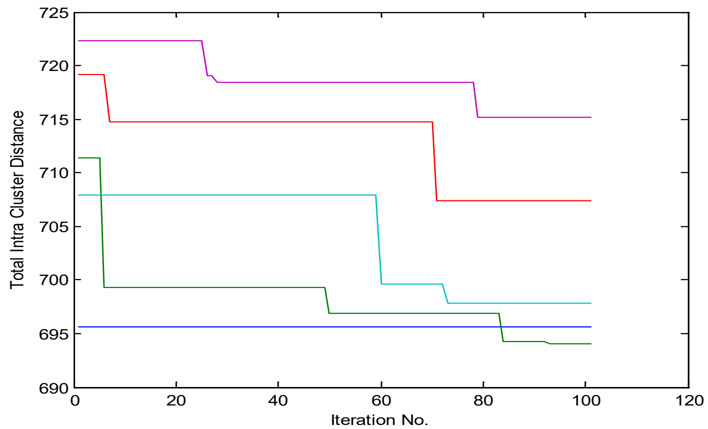

Figure 14. Convergence characteristics in 2nd stage for multidomain MMDE

Table 13

Centroid values by Multidomain MMDE

				Centroids of MMDE			
C1	16.0000	1.5165	13.4754	3.3530	2.4072	74.6342	0.0100
C2	201.3622	1.5122	14.7074	0.1029	1.2528	72.3216	0.1859
C3	165.4855	1.5189	12.7370	2.3479	2.1774	71.8032	0.7419
C4	48.0214	1.5246	11.9324	4.4900	1.1781	72.9279	0.7290
C5	88.8809	1.5116	13.4721	3.3903	1.0875	72.9210	0.3255
C6	127.1936	1.5134	13.9751	3.8544	1.4775	73.6876	0.2323

Table 13 (*Continued)*

	Centroids of MMDE		
C1	8.7993	0.0894	0.2050
C2	8.6580	1.3473	0.0031
C3	7.7070	0.2396	0.0068
C4	9.8281	0.0987	0.0876
C5	7.9812	0.0100	0.1157
C6	9.0625	0.0100	0.1454

Comparative Study of MMDE with K-Means

In practice it has been observed that K-means algorithm is very effective and useful along with having the dominance in the utilization. In fact it is one of the best algorithms in terms of computational cost and efficiency.

Comparative performance between Multi-Domain MMDE and K-Means over all the three different data sets are shown in Table 14 to Table 16. For each data set 5 independent trials had been applied. It can be understood from the outcomes that the problems with K-Means algorithm are twofold.

First it may not deliver the optimal performances, second, there is high level of variations in the performances over trials which is a really serious issue from the practical point of view. This happens because of sensitivity of K-Means algorithm towards initialization. Whereas the proposed method Multi-domain MMDE has delivered not only better performance because of exploration but also the variation level is marginal.

Table 14

Comparative Performance of MMDE and K-means for Wine Data

WineData	**Multi-Domain MMDE Samples**		**K-Means K means Samples**	
Trial	**Correctly clustered**	**Wrongly Clustered**	**Correctly clustered**	**Wrongly Clustered**
1	125	53	125	53
2	125	53	120	58
3	125	53	120	58
4	125	53	120	58
5	125	53	120	58
Mean	**125**	**53**	**123.75**	**54.28**
Efficiency	**70.22**		**67.98**	

Table 15

Comparative Performance of MMDE and K-means for Iris Data

Iris Data	**Multi-Domain**		**K-Means**	
	MMDE Samples		**K means Samples**	
Trial	**Correctly clustered**	**Wrongly Clustered**	**Correctly clustered**	**Wrongly Clustered**
1	135	15	134	16
2	134	16	134	16
3	137	13	100	50
4	133	17	134	16
5	134	16	100	50
Mean	**134.6**	**15.4**	**120.4**	**29.6**
Efficiency	**89.73**		**80.27**	

Table 16

Comparative Performance of MMDE and K-means for Glass Data

Glass Data	**Multi-Domain**		**K-Means**	
	MMDE Samples		**K means Samples**	
Trial	**Correctly clustered**	**Wrongly Clustered**	**Correctly clustered**	**Wrongly Clustered**
1	188	26	187	27
2	188	26	187	27
3	189	25	187	27
4	190	24	187	26
5	191	33	187	27
Mean	**187.2**	**26.8**	**187**	**26.8**
Efficiency	**87.48**		**87.38**	

CONCLUSION

In this paper, a Modified Mutation Strategy for Differential Evolution (MMDE) has been proposed to facilitate the clustering requirement of data. This modification increases the convergence rate and delivers satisfactory cluster efficiency. To increase the level of exploration, two stage based a multimodal structure has also been proposed. With this structure, the bias variation sensitivity of cluster activity decreases. Number of benchmarks had been tested that had the number of clusters from 2 to 6 to ensure the algorithms generalized capability. Proposed solution has outperformed the Conventional form of DE as well as Dynamic weighted form of PSO. Proposed work had been evaluated only using datasets of UCI Repository, further it could be applied on application oriented dataset to evaluate performance.

ACKNOWLEDGEMENT

SP would like to offer my special thanks to Dr. Anandhi Rajamani Jayadharmarajan, Professor and Head, Department of Information Science and Engineering, New Horizon College of Engineering, Bengaluru, my research supervisor, for her patient guidance, enthusiastic encouragement and useful critiques of this research work. SP would also like to thank the Principal, The Oxford College of Engineering, Bengaluru, for his advice and assistance in keeping my progress on schedule. SP would like to thank UCI machine learning repository for providing datasets. Finally, SP wishes to thank her husband and parents for their support and encouragement throughout her research work. We gratefully thank the Visvesvaraya Technological University, Jnana Sangama, Belgavi for providing opportunity to conduct this research work.

REFERENCES

Das, S., Abraham, A., & Konar, A. (2008). Automatic clustering using an improved differential evolution algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 38*(1), 218-237.

Dua, D., & Graff, C. (2019). *UCI machine learning repository*. Irvine, California: University of California, School of Information and Computer Science.

Gupta, Y., & Saini, A. (2018). A new swarm-based efficient data clustering approach using KHM and fuzzy logic. *Soft Computing, 23*(1), 145-162.

Jiau, H. C., Su, Y. J., Lin, Y. M., & Tsai, S. R. (2006). MPM: A hierarchical clustering algorithm using matrix partitioning method for non-numeric data. *Journal of Intelligent Information Systems, 26*(2), 185-207.

Kuo, R. J., & Zulvia, F. E. (2019). An improved differential evolution with cluster decomposition algorithm for automatic clustering. *Soft Computing, 23*(18), 8957-8973.

Liu, X., Yuan, J., & Zhao, H. (2018). Efficient and intelligent density and delta-distance clustering algorithm. *Arabian Journal for Science and Engineering, 43*(12), 7177-7187.

Nerurkar, P., Pavate, A., Shah, M., & Jacob, S. (2018). Performance of internal cluster validations measures for evolutionary clustering. *Computing, Communication and Signal Processing, 810*, 305-312.

Tran, D. C., Wu, Z., & Deng, C. (2015). An improved approach of particle swarm optimization and application in data clustering. *Intelligent Data Analysis, 19*(5), 1049-1070.

Wang, H., Zuo, L. L., Liu, J., Yi, W. J., & Niu, B. (2018). Ensemble particle swarm optimization and differential evolution with alternative mutation method. *Natural Computing, 18*(71), 1-14.

Yi, H., Niu, Z., Zhang, F., Li, X., & Wang, Y. (2018). Robust recommendation algorithm based on kernel principal component analysis and fuzzy C-means clustering. *Wuhan University Journal of Natural Sciences, 23*(2), 111-119.

Zhu, Y., Zhang, M., & Shi, F. (2018). Application of algorithm CARDBK in document clustering. *Wuhan University Journal of Natural Sciences, 23*(6), 514-524.