Clustering with modified mutation strategy in Differential Evolution

Seema Patil*, R. J Anandhi **

** Department of Computer Science & Engineering, The Oxford College of Engineering*
*** Department of Information Science & Engineering, New Horizon College of Engineering*

Abstract- In this paper, a clustering approach based on modified mutation strategy in the Differential Evolution has been proposed. The objectives of modification are to achieve high rate of convergence and to obtain better cluster efficiency. The proposed form of modification has been applied on probabilistic environment to define the differential vector through randomly selected members and the best solution has been obtained. Over number of benchmark dataset, clustering efficiency have been estimated and compared with Conventional Differential Evolution as well as Particle Swarm Optimization. The proposed solution has delivered the superior and consistent performance over the considered benchmark.

Index Terms-Clustering, Convergence, Differential evolution, Mutation, Particle swarm optimization

I. INTRODUCTION

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

To perform this, knowledge extraction is done with the help of data mining techniques such as classification and clustering. The important task of combining various population or data points into clusters is clustering which performs similarity of points. It is one of iterative process of discovery of 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 [1]. Clustering defines a new swarm intelligence (SI) for partitioning any datasets into an optimal number of groups through one run of optimization. SI is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena 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, or swarms of bees, and even human social behavior, from which the idea is emerged. Data clustering using PSO can be used to find the centroids of a user specified number of clusters. For automatic clustering of large unlabeled data sets, Differential Evolution (DE) is used. [2]

This work proposed the method for clustering, based on differential evolution. Even though DE is very efficient, but sometimes it suffers from the issue of slow convergence and difficulties in achieving the 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, deliver the fast change in the solution and results in faster convergence. The multi-culture approach helps in exploration of new and efficient solution. Gathering and selection of solution from different environments will maintain the diversity in the population.

II. RELATED WORK

The author Gupta [3] et al., has 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. The author Pranav [4] et al., has achieved the global optima on clustering by making use of two validation indices criteria. These indices were simple and robust against other outliers and shown best clustering which has lower computation cost and parallel execution and faster convergence. The author Wang [5] et al., combines 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 are performed throughout the optimization process to get the good results. This approach gives the best result compared to inertia weight PSO and comprehensive learning PSO and basic DE. The author Zhu et al., [6] has discussed complications associated with K-means clustering algorithm and centroid all rank distance concept has been presented. To overcome the difficulties associated with density and delta-distance clustering (DDC) when data derived from the two indicators are large, an efficient and intelligent DDC algorithm has been discussed by author Liu et al [7]. A robust recommendation algorithm based on kernel principal component analysis and fuzzy c-means clustering has been presented by author Huawei et al., [8]. The author has presented a variation of differential evolution (DE) algorithm to solve an automatic clustering problem [9]. The author [10] describes 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. Based on matrix partitioning a hierarchical clustering algorithm has been presented in [11].

III. PROPOSED WORK

A. Modified Mutated DE (MMDE)

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 generate the faster change, while the random member-based selection tries to prevent from suboptimal convergence. The pseudo code for applied mutation strategy has been shown below.

- Define the Threshold value (*Thr*)
- 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*

B. 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 exploit to form a better community to search the solution space efficiently. This approach is inspired very much by present human society, where at fundamental level two things happen (i) the independent existence of a number of separate population, and they get their progress under the same environment up to a certain period of time. (ii) with respect to objectives, a number of individuals are selected from the different population and form a new population to achieve the objectives. Rather than working under monoculture formed by one population as in conventional PSO, multiculture environment has been proposed, where a number of different environments created by a different set of population independently. Each population has evolved socially, independently to generate the multiculture and later among all, best individuals are selected to finish the task. This is a dual stage process where first stage finds some potential solution discovered from different regions of solution space, and later in the second phase, each individual contributes more efficiently to find a global solution. Even with the small size of the population, the proposed method has achieved better quality solution with the very high value of consistency.

In the working principle of MMDE, population (POP) are the initial random population, which is evolved by the DE process individually and independently for a fewer number of 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 surrounding, each NPOP has different characteristics. Through the fitness-based selection process, among all members from all NPOP, better members are selected to form a new population (SPOP), which has the same size as initial POP. In SPOP, there are a number of good candidates, which are different and have higher fitness value, hence the high level of diversity exists. Finally, over SPOP, MMDE has been applied till terminating criteria has not meet, to obtain the Final Population (FPOP).

IV. EXPERIMENTAL RESULTS AND ANALYSIS

For the data set namely "Wine data"," Iris", and "Glass" data set which are available in UCI repository[12] have been considered to analyze the work. In the first part, only the MMDE has been applied and performances have been obtained for 5 independent trials. Comparison has been made with conventional DE(CDE) and dynamic weighted PSO(DYPSO). For all the cases, the size of population has been considered as 100, mutation rate and crossover rate as 0.4 and 0.5. The allowed number of iterations were 600.The performances have been represented in terms of correctly placed data samples in the clusters, number of data samples placed wrongly, cluster efficiency and total intra cluster distance value. In second part, multidomain based experiment has been included with MMDE and performances have been estimated over "Glass" data set. Experimental process has been developed in the MATLAB environment.

A. Dataset: Wine Data

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

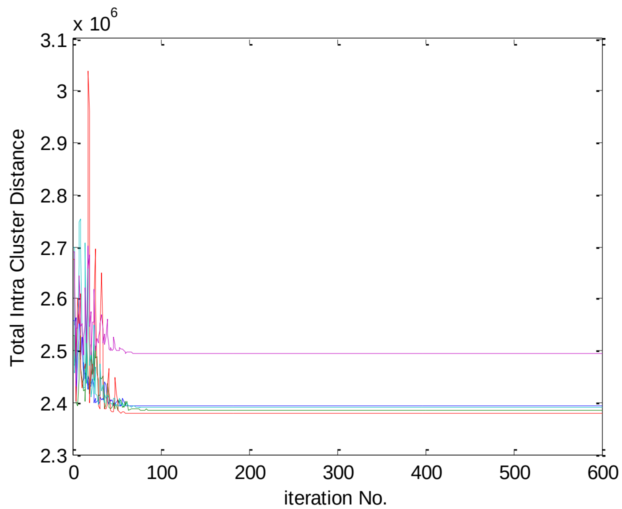


Fig.1 DYPSO based convergence in 5 trials for winedata set

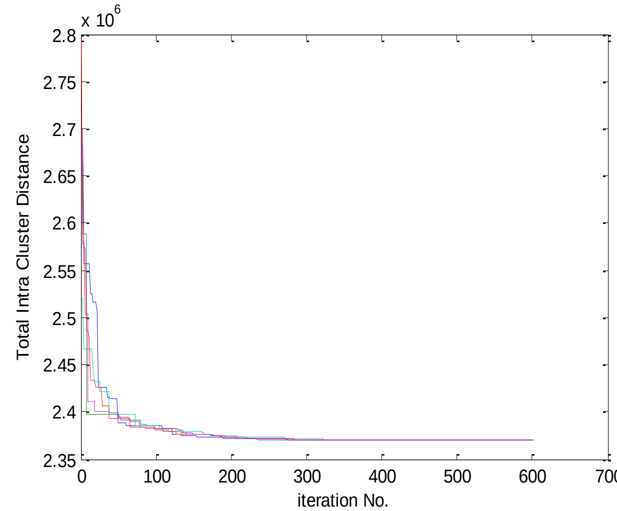


Fig.2 CDE based convergence in 5 trials for winedata set

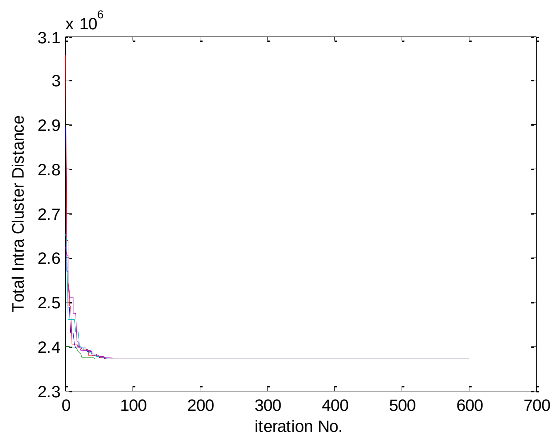


Fig.3: MMDE based convergence in 5 trials for winedata set

Table1: Mean Performance over 5 trials by different algorithm over winedata 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

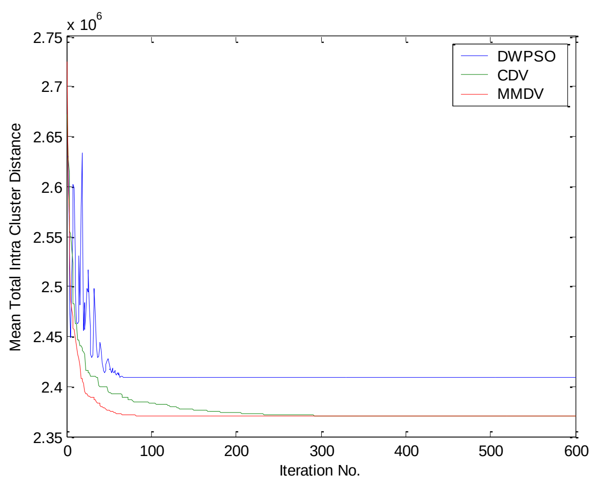
Table 2: Centroid position for winedata

C1	3.0351	3.0067	3.0065	3.0541	3.2816	3.0057	3.0043	3.0108	3.0041	3.0154	3.0024	3.0067	4.9797
C2	3.0375	3.0051	3.0065	3.0462	3.2867	3.0078	3.0081	3.0008	3.0051	3.0154	3.0029	3.0084	6.2486
C3	3.0339	3.0067	3.0062	3.0565	3.2508	3.0057	3.0048	3.0010	3.0040	3.0111	3.0024	3.0067	4.2455

The performances obtained under 5 independent trials by different algorithms have been shown in Table1.It can be observed that all the three algorithms have nearly the same performances, while there is little more distance measure appeared for the DYPSO. The obtained centroid value by MMDE for 1st trail have been shown in Table2.The convergence characteristics for DYPSO, CDE and MMDE have been shown in Fig.1 to Fig.3.To get the relative convergence speed, Fig.4 has plotted the mean convergence characteristics. Proposed MMDE has shown the fastest rate of convergence while DYPSO was the poorest.

B Dataset: IRIS Data

Contain total 150 data set and each data has 4 attributes. Three different global clusters exist in dataset. The convergence performances of DYPSO, CDE and MMDE have been shown in Fig. 5 to Fig.7, while the statistical performances have been shown in Table 3 to Table 5. It can be observed that MMDE has shown very consistent performance in all trials and in Fig.8 comparative convergence has been shown. The obtained best value of centroid has been shown in Table 6.



Fig,4: Mean convergence comparison for Iris data set

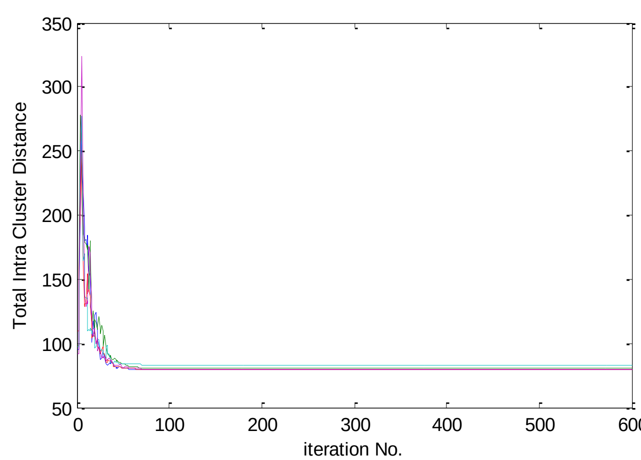


Fig.5: DYPSO based convergence in 5 trials for Iris data set

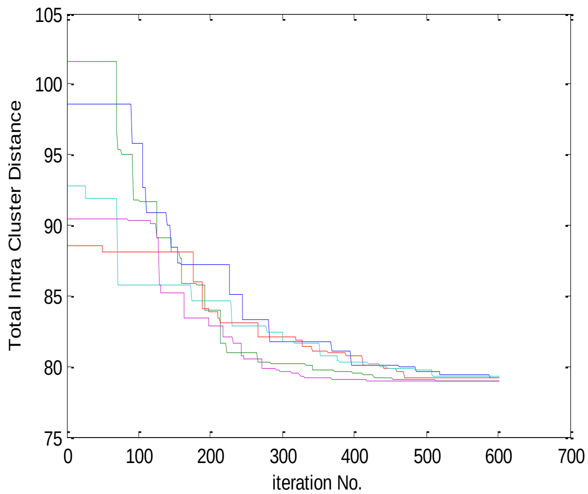


Fig.6: CDE based convergence in 5 trials for Iris data set

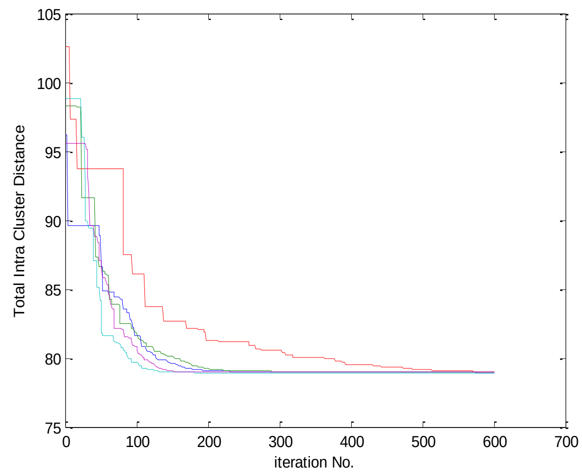


Fig.7: MMDE based convergence in 5 trials for Iris data set

Table 3: DYPSO performance over Iris data

Trial No. IRIS(PSO)	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**

Table 4: CDE performance over Iris data

Trial No. IRIS(CDV)	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**

Table 5: MMDE performance over Iris data

Trial No. IRIS(MMDV)	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**

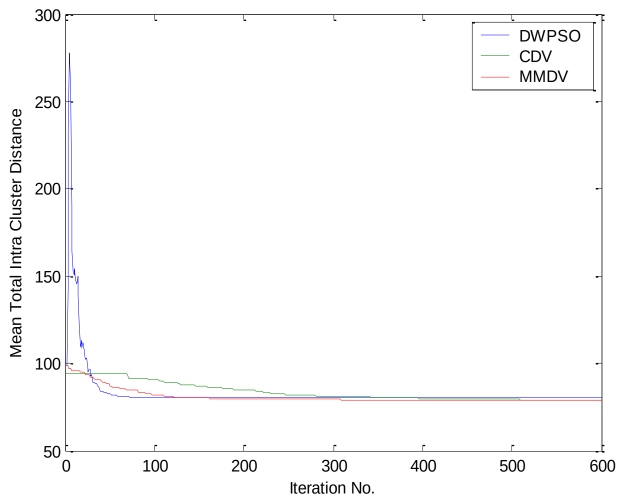


Fig.8: Mean convergence comparison for Iris data set

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

Table 6: Centroids value for Iris data set

C. Dataset: Glass Data

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

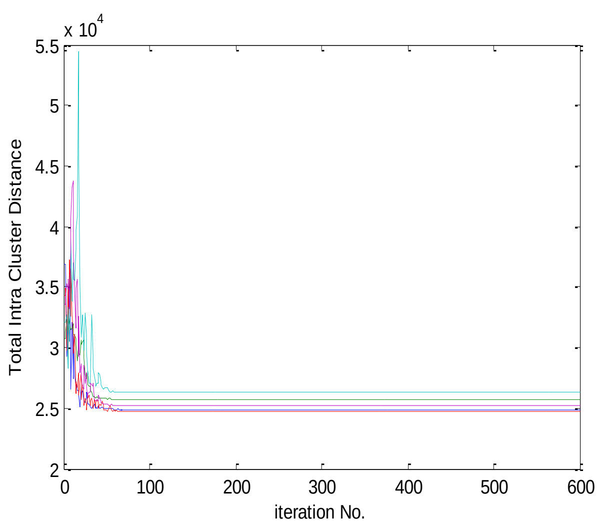


Fig.9: DYPSO based convergence in 5 trials for Glass data set

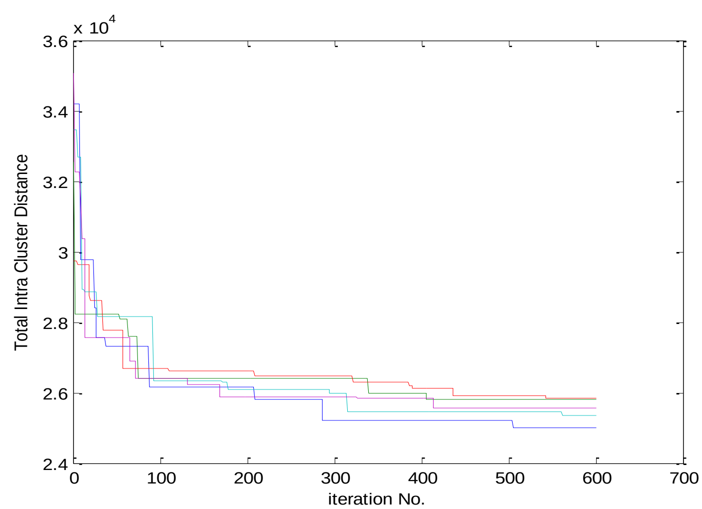


Fig.10: CDE based convergence in 5 trials for Glass data set

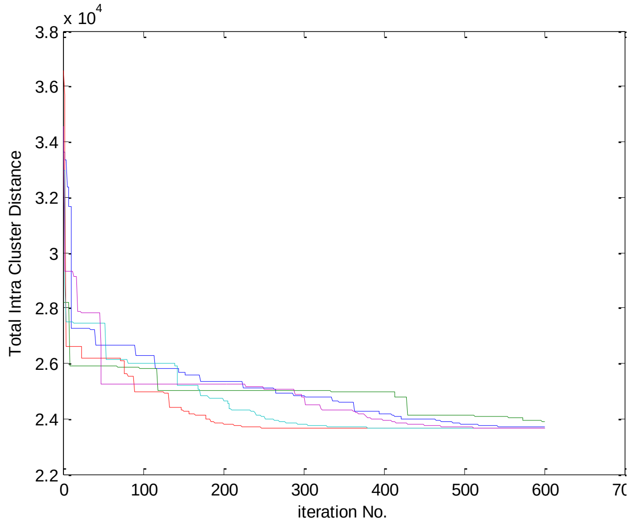


Fig.11: MMDE based convergence in 5 trials for Glass data set

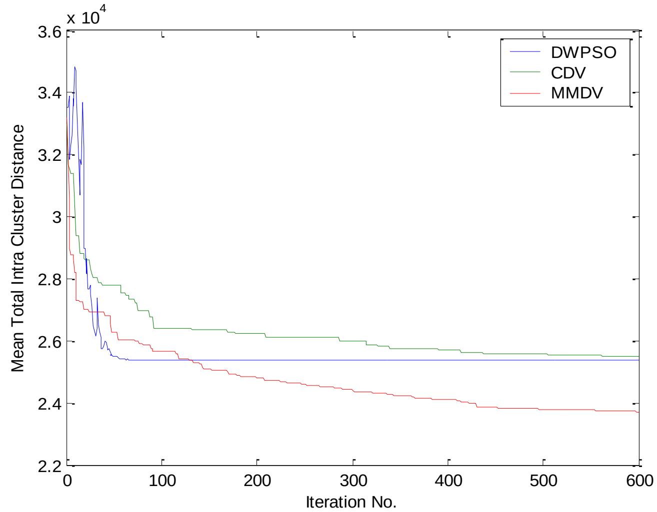


Fig.12: Mean convergence comparison for Glass data set

Table 7: DYPSO performance over Glass data

Trial No. Glass (PSO)	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
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**

Table 8: CDE performance over Glass data

Trial No. Glass (CDE)	Correctly clustered data samples	Wrongly clustered data samples	Clustered efficiency	Total Intra Cluster Distance value
1	183	31	85.51	2.4990 e+004
2	189	25	88.32	2.5797 e+004
3	178	36	83.18	2.5850e+004
4	184	30	85.98	2.5368 e+004
5	188	26	87.85	2.5546 e+004
Mean	**184.4000**	**29.6000**	**86.17**	**2.5510e+004**

Table9: MMDE performance over Glass data

Trial No. Glass (MMDE)	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**

Table10: Centroids value for Glass data set

C1	166.0782	2.4471	13.7061	3.5266	2.2563	73.3031	2.4611	10.7421	-0.1976	0.5747
C2	198.4844	2.5638	16.2827	3.2212	2.7751	73.5565	1.7972	9.9803	1.6024	-0.1853
C3	54.2369	2.1344	14.2542	4.4666	1.9043	72.6730	1.0457	9.7003	1.4352	0.2335
C4	18.5031	2.1863	13.2582	4.4278	1.5191	74.4194	1.3567	10.2181	0.4565	10.1096
C5	129.9205	0.8875	13.9521	4.3390	2.7228	75.5818	0.9168	8.7067	1.4468	1.4522
C6	91.0957	2.8459	14.1901	3.6017	2.9122	72.2789	0.9257	10.0617	0.7071	1.1787

For the Glass data set the obtained convergence characteristics have been shown in Fig.9 to Fig.11. Comparative mean convergence has been shown in Fig.12. It can be observed that, in spite of more number of clusters, superior convergence has appeared. The obtained statistical performance has been shown in Table7 to Table9. For MMDE, maximum cluster efficiency has been obtained. The obtained best centroid value has also been shown in Table10.

D. Multidomain based MMDE

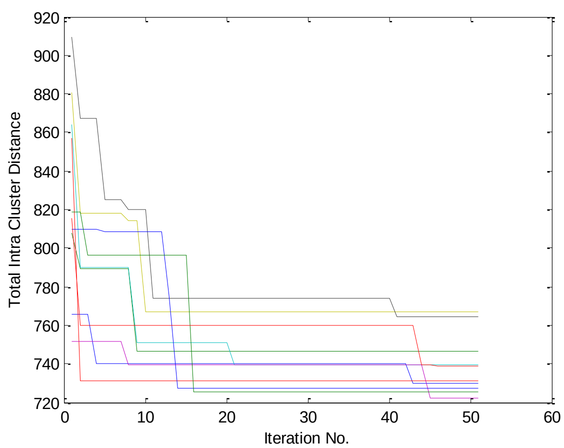


Fig.13 Convergence characteristics in 1st Stage for multidomain MMDE

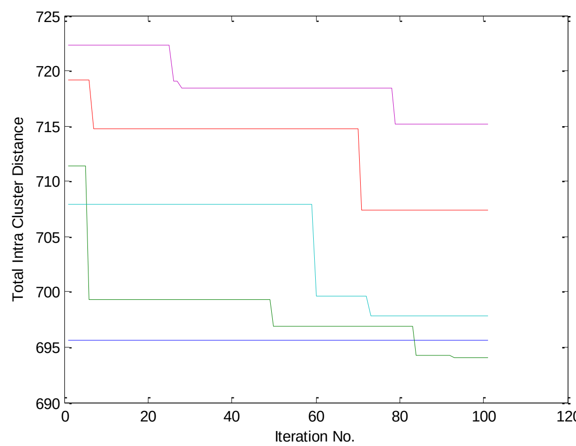


Fig.14 Convergence characteristics in 2nd stage for multidomain MMDE

Convergence characteristics over Glass data set for multidomain MMDE has been shown in Fig.13, for the 1st stage and in Fig.14 for the 2nd stage. The obtained performances have been shown in Table11. It can be observed that maximum efficiency 87.48% has been obtained. The corresponding centroid value has also been shown in Table 12.

Table11: Multidomain MMDE performance over Glass data

Trial No. (MMDE) GLASS	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
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)**

Table12: Centroid values by Multidomain MMDE

C1	16.0000	1.5165	13.4754	3.3530	2.4072	74.6342	0.0100	8.7993	0.0894	0.2050
C2	201.3622	1.5122	14.7074	0.1029	1.2528	72.3216	0.1859	8.6580	1.3473	0.0031
C3	165.4855	1.5189	12.7370	2.3479	2.1774	71.8032	0.7419	7.7070	0.2396	0.0068
C4	48.0214	1.5246	11.9324	4.4900	1.1781	72.9279	0.7290	9.8281	0.0987	0.0876
C5	88.8809	1.5116	13.4721	3.3903	1.0875	72.9210	0.3255	7.9812	0.0100	0.1157
C6	127.1936	1.5134	13.9751	3.8544	1.4775	73.6876	0.2323	9.0625	0.0100	0.1454

E. Comparative study of MMDE with K-Means

Comparative performance between Multi-Domain MMDE and K-Means over all the three different data sets have been shown in Table13-15. For each data set 5 independent trials have been applied. It can be observed with 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 trails which is really a 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 variation level is very less.

Table 13: Comparative Performance of MMDE and K-means for Wine Data

WineData	**Multi-Domain**		**K-Means**	
Trial	**MMDE Samples**		**K means Samples**	
	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 14: Comparative Performance of MMDE and K-means for Iris Data

Iris Data	Multi-Domain		K-Means	
Trial	MMDE Samples		K means Samples	
	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 15: Comparative Performance of MMDE and K-means for Glass Data

Glass Data	Multi-Domain		K-Means	
Trial	MMDE Samples		K means Samples	
	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**	

V. CONCLUSION

In this paper, a modified mutation strategy for differential evolution has been proposed to facilitate the clustering requirement of data. This modification increases the convergence rate and deliver the cluster efficiency up to the mark. 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 decreased. Number of benchmarks have been tested which had the number of clusters from 2 to 6 to ensure the generalize capability. Proposed solution has outperformed the conventional form of DE as well as dynamic weighted form of PSO. Proposed work has been evaluated only using datasets of UCI Repository, further it can be applied on application oriented dataset to evaluate performance.

ACKNOWLEDGMENT

I would like to offer my special thanks to Dr. R J Anandhi, 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. I would also like to thank Dr. R V Praveena Gowda, Principal, The Oxford College of Engineering, Bengaluru, for his advice and assistance in keeping my progress on schedule. Finally, I wish to thank my husband and parents for their support and encouragement throughout my research work.

References

[1] "Nuria Gómez Blas, Octavio López Tolic, "Clustering using Particle Swarm Optimization", International Journal Information Theories and Applications. 23, pp24-33, 2016

[2] Swagatam Das, Ajith Abraham, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 38, NO. 1, pp 218-237, JANUARY 2008

[3] Yogesh Gupta, Ashish Saini, "A new swarm-based efficient data clustering approach using KHM and fuzzy-logic", Soft Computing, Springer, pp 145-162, 2019.

[4] Pranav Nerurkar, Aruna Pavate, Mansi Shah and Samuel Jacob, "Performance of Internal Cluster Validations Measures for Evolutionary Clustering", Computing, Communication and Signal Processing, Springer pp 305-312 2019

[5] H. Wang, L. L. Zuo, J. Liu, W. J. Yi, B. Niu1, "Ensemble particle swarm optimization and differential evolution with alternative mutation method", Natural Computing, Springer Nature pp 1-14,2018

[6] Yehang Zhu,Mingjie Zhang,Feng Shi, "Application of Algorithm CARDBK in Document Clustering", Wuhan University Journal of Natural Sciences,December 2018, Volume 23, Issue 6, pp 514–524

[7] Xuejuan Liu,Jiabin Yuan,Hanchi Zhao, "Efficient and Intelligent Density and Delta-Distance Clustering Algorithm", Arabian Journal for Science and Engineering,December 2018, Volume 43, Issue 12, pp 7177–7187

[8] Huawei Yi "Robust Recommendation Algorithm Based on Kernel Principal Component Analysis and Fuzzy C-means Clustering", Wuhan University Journal of Natural Sciences,April 2018, Volume 23, Issue 2, pp 111–119.

[9] Kuo, R.J. & Zulvia, F.E. "An improved differential evolution with cluster decomposition algorithm for automatic clustering", Soft Computing, Springer, pp 1–17,2018.

[10] Dang Cong Trana,c,, Zhijian Wua and Changshou Dengb, "An improved approach of particle swarm optimization and application in data clustering", Intelligent Data Analysis 2015 pp 1049–1070

[11] Hewijin Christine Jiau,"MPM: a hierarchical clustering algorithm using matrix partitioning method for non-numeric data", Journal of Intelligent Information Systems,May 2006, Volume 26, Issue 3, pp 303–306.

[12] https://archive.ics.uci.edu/ml/index.php