Modified Evolutionary Fuzzy Clustering

Bhavana Devi
Research Scholar
Dept. of CSE,
Rama University, Kanpur, India

Vivek Srivastava
Associate Professor
Dept. of CSE,
Rama University, Kanpur, India

Abstract: **In this paper, a novel modified evolutionary fuzzy clustering has been proposed. This technique exploits parameters such as Minkowski distance, Xie Beni index, and classification entropy. These parameters turn to account to polish the performance measure by incorporating shape of the cluster, class compactness and partitioning quality respectively in standard Fuzzy C-Means clustering. Moreover, we also employed an evolutionary strategy for optimizing fuzzy partition matrix. This avenue is examined over nine real world bench-mark problems of various domains. We ascertained that each problem sustains optimal result as compared to the conventional ones.**

Keywords: *Clustering, Fuzzy C-Means clustering, Evolutionary strategy, Multi Objective Evolutionary Fuzzy clustering.*

I. INTRODUCTION

Clustering is the task of grouping several objects into a number of clusters such that objects in the same group are more similar in some context than the other ones. Clustering is broadly divided into hard and soft clustering. In hard clustering, data is divided into different clusters, where each data point belongs to exactly one cluster. This harsh requirement makes hard clustering not likely to accept in the case where grouping decision is not definite. While in soft clustering, each object has some degree of belonging i.e. one object can be the part of two or more clusters. Fuzzy C-Means clustering [1] [2] is widely used soft clustering technique. However, it may fail in the cases where high dimensional datasets [3] [4] and large no of prototype exists, when the objects are only slightly different from each other. Another remarkable drawback of Fuzzy C-Means clustering is generation of different partition matrix on different runs of same algorithm. Therefore, in many researches; concepts of evolutionary techniques have been incorporated for obtaining optimal fuzzy partition matrix. Henceforth, various kinds of evolutionary algorithms have been proposed in various literatures [5] [10] for Fuzzy C-Means clustering. Three major concepts of evolutionary computation involve evolutionary algorithm [6] [7], evolutionary strategy and evolutionary programming [3] [8].

In this paper, we introduce a novel evolutionary strategy for Fuzzy C-Means clustering. In evolutionary strategy, the output of the previous stage acts as the input in the next stage. Here, it is being employed to improve the partitioning quality of the fuzzy clusters. In conventional Fuzzy C-Means clustering, Euclidian distance is often used as distance measure. Euclidean distance always tends to generate spherical clusters, which will be inappropriate where the natural clusters for the data are not spherical. Therefore, we have employed Minkowski distance measure because it is a generalized distance measure involving a parameter ‘p’ for generalization.

Divergence of validity indices [11] are used in conventional Fuzzy C-Means clustering. These validity criteria determine the clustering tendency of a set of data or used to measure the fitness of the partition. Validity indices are broadly divided in two types, external and internal validation. In external validation, we evaluate the results of the clustering algorithm based on a pre specified structure, which is imposed on a dataset. Internal validations are used to measure the goodness of a clustering structure without respect to external information. Here, we employed classification entropy and Xie Beni index. Entropy belongs to external validation indices. It involves only membership function not dataset. We apply classification entropy because of its heuristic measure of the rationale underlying its formulation. This index is a scalar measure of the amount of fuzziness in a given partition matrix. Xie Beni is employed here because it deals with the membership function as well as dataset. It focuses on the compactness as well as the separation. A good partition provides a small value for the compactness and well separated c_i will produce high value for separation [12].

This paper presents a novel modified evolutionary fuzzy clustering. This technique utilize parameters such as Minkowski distance [13], Xie Beni index [14] [15], and classification entropy [16] [17] along with evolutionary strategy. These parameters improve the performance measure by incorporating shape of the cluster, class compactness and partitioning quality respectively.

The remainder of this paper is organized as follows: Section II presents Fuzzy c-means clustering and its variants, whereas Section III presents proposed Modified evolutionary fuzzy clustering. Section IV presents Experimental results and discussion. Finally, Section V involves conclusion.

II. FUZZY C-MEANS CLUSTERING AND ITS VARIANTS

Fuzzy C-Means clustering was developed by Dunn in 1973 [19] and improved by Bezdek [5] in 1981. It is based on the concept of fuzzy c-partition which is the generalization of crisp partition. For a crisp partition, the values belong to either 0 or 1, i.e. an object can belong to one and only one cluster. For a fuzzy partition, the object belongs to multiple clusters with varying degrees of belongingness. Fuzzy C-Means clustering is a generalization of K-Means clustering. The main aim of fuzzy clustering is to classify N objects into C clusters. It is based on the minimization of the objective function given as

$$J_m(U,c) = \sum_{j=1}^{C}\sum_{i=1}^{N} u_{ji}^{m}\,||x_i - c_j||^2 \qquad (1)$$

Where N is the number of objects in the measured dataset, C is the number of clusters, u_{ij} is the degree of membership, m (fuzzifier index) is any real number greater than 1, x_i is the *i*th object of the d dimensional measured data and c_j is the center of the *j*th cluster with d dimension. Here for $1 \le i \le N$ and $1 \le j \le C$, we have $0 \le u_{ij} \le 1$.

Basic Fuzzy C-Means clustering is composed of the following steps:

1. Given a fixed number of clusters C and select a proper termination condition, ϵ
2. Initialize fuzzy partition (membership) matrix $U = [u_{ij}]$ matrix, U^0
3. Determine the cluster's center c_j ($1 \le j \le C$) by:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \qquad (2)$$

4. Update the fuzzy partition matrix U^k as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}} \qquad (3)$$

5. If $||U^{k+1} - U^k|| < \epsilon$ then STOP, otherwise go to step 3.

Here, $||U^{k+1} - U^K||$ is defined to be absolute value of $||u_{ij}^{new} - u_{ij}^{old}||$ and $U = (u_{ij})$ where u_{ij} represent the degree of object x_i belonging to the *i*th cluster.

In many researches, it has been observed that evolutionary algorithms are successfully combined with Fuzzy C-Means Clustering for yielding better results. In research work [20], an enhanced quantum inspired evolutionary fuzzy clustering has been proposed to compute global optimal value of fuzzifier, number of clusters and initial cluster centroid. Quantum computing concept is combined with fuzzy clustering to develop the distinct values of these parameters in several generations. In research [21], this paper evolves an evolutionary approach to automatically determine the optimal number and location of prototypes for the well known Fuzzy C-Means clustering algorithm. This method is based on clustering through genetic algorithm and employ context-sensitive genetic operators to globally explore the search space. In paper [22], the author considers the problem of finding clusters in a high dimensional data. Evolutionary approach integrates Fuzzy C-Means clustering and feature selection. Evolutionary approach reduces the dimensionality of the space and feature selection improves the quality of the partitions that are being generated. In paper [18], the evolutionary algorithm has been evolved for determining the best distance for given data. The standard of goodness is defined in terms of the performance of the Fuzzy C-Means clustering algorithm.

Hence, it has been widely recognized that incorporation of evolutionary algorithms with fuzzy clustering results into performance improvement. Therefore, we have proposed a novel multi-objective evolutionary fuzzy clustering that is free from the major deficiency of conventional fuzzy c means clustering. This novel proposal is presented in next section.

III. MODIFIED EVOLUTIONARY FUZZY CLUSTERING

There are some flaws in fuzzy clustering that makes it less effective (inappropriate) in finding a high quality fuzzy partition matrix and discovering the arbitrary shape of the cluster. In conventional Fuzzy C-Means clustering, Euclidian distance is often used as distance measure. Euclidean distance always tends to generate spherical clusters, which will be inappropriate where the natural clusters for the data are not spherical. It has been widely observed in many researches [23] [18], that the shapes of clusters are not always spherical for real world data. Therefore, we have employed Minkowski distance measure because it is a generalized distance measure involving a parameter *'p'* for generalization. Euclidean distance is an instance of Minkowski distance at generalization parameter *p=2.*

$$d(x,c) = \left(\sum_{i=1}^{n} |x_i - c_i|^p\right)^{\frac{1}{p}} \quad p > 0 \qquad (4)$$

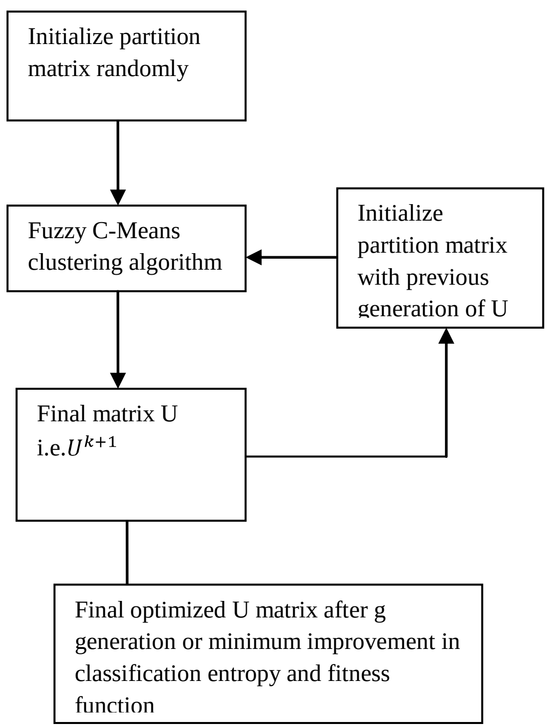

Figure.1 Multi objective Evolutionary fuzzy clustering strategy

Another remarkable drawback of Fuzzy C-Means clustering is generation of different partition matrix on different runs of same algorithm. To overcome this problem, concept of evolutionary strategy has been incorporated for obtaining optimal fuzzy partition matrix. Evolutionary strategy is based on the approach that the outcome of the previous step acts as the input in the next step. This is applied in fuzzy partition matrix (U) employed in the Fuzzy C-Means clustering, it is randomly initialized in first iteration and in next iteration, the best value of the partition matrix is applied as an input to the next generation.

Modified Evolutionary fuzzy clustering implies multiple objectives that are being optimized simultaneously. Here, multiple objectives employ class compactness, partitioning quality and shape of the cluster.

These objectives are fulfilled by Xie Beni Index, classification entropy and minkowski parameter respectively. In researches, it has been found that different validity criteria are used to address different properties of fuzzy clustering to improve results. These validity indices measure the goodness of the fuzzy partition matrix as well as polish the fuzzy partitions of the cluster. Validity indices are widely divided in two types, external and internal validation. In external validation, we assess the results of the clustering algorithm based on a pre specified order, which is enforced on a dataset. Internal validations are acclimated to measure the goodness of a clustering structure without respect to external information. Here we employed classification entropy and Xie Beni index. Entropy belongs to external validation indices. It involves only membership function not dataset. We apply classification entropy because of its heuristic measure of the hypothesis underlying its formulation. Xie Beni is employed here because it deals with the membership function as well as dataset. It endeavors on the compactness as well as the separation. A good partition provides a small value for the compactness and well separated c_i will produce high value for separation [12].

Modified evolutionary fuzzy clustering is composed of the following steps where f symbolizes to fitness function. €, £, ∈ are pre-defined threshold values.

Algorithm:

Input:x_i $(dataset)$, C (number of clusters), q (user defined parameter).

Method:

1. Initialize fuzzy partition matrix $U^0 \leftarrow [u_{ij}]_{C*N}$ $randomly$ and $p \leftarrow 1$
2. Repeat step 2 to 6 while$(p < q)$

$$d(x,c) = \left(\sum_{i=1}^{n} |x_i - c_i|^{p}\right)^{\frac{1}{p}} \quad p > 0 \quad (5)$$

3. Calculate center $c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$ (6)

4. Update partition matrix U^k

$$[u_{ij}]_{C\times N} = \frac{1}{\sum_{k=1}^{C}\left(\frac{||x_i - c_j||}{||x_i - c_k||}\right)^{\frac{2}{m-1}}} \quad (7)$$

5. If $||U^{k+1} - U^k|| \leq \epsilon$ then go to step 6
Else go to step 3.
.

6. $Xiebeni = \frac{\sum_{j=1}^{C}\sum_{i=1}^{N} u_{ji}^{m}\,||x_i - c_j||^2}{N\times m}$ (8)

7. $f = \frac{1}{Xiebeni}$

8. $classification\ entropy(CA^k) = \sum_{j=1}^{C}\sum_{i=1}^{N}\frac{U_{ji}^{k+1}.\log\left(U_{ji}^{k+1}\right)}{N}$ (9)

9. $U^0 \leftarrow U^{k+1}$

10. Repeat step 2 to 8 until $||f^{new} - f^{old}|| < \in$ and $||CA^{k+1} - CA^k|| \leq £$ or U^g

11. $p \leftarrow p + 1$. Go to step 2

Output: U^g $('g'\ no\ of\ generation)$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. *Datasets*

We have employed 9 benchmark problems retrieved from UCI Machine learning Repository [24] whose description is mentioned below:

i. Breast cancer: This problem contains 569 instances, 32 attributes and two class predictions, Benign and Malignant. 357 data points for Benign and 212 for Malignant.
ii. Wine dataset: This is a three class problem having 178 instances and 13 attributes. Class 1 has 59 instances, class 2 has 70 and class 3 has 48 instances.
iii. Iris dataset: This is also a three class problem having 150 instances and each class has 50 instances.
iv. User knowledge modeling: The problem contains 403 instances, 5 attributes having four classes. Class 1 contains 50 instances, class 2 has 128, class 3 has 121 and class 4 has 101 instances.
v. Blood transfusion service center: This problem has 748 data points, 5 attributes and two classes.
vi. Fertility dataset: This is also a 2 class problem having 100 instances and 10 attributes. Class1 (normal-88) and class2 (altered-12).
vii. SPECTF heart dataset: This dataset problem has 267 instances, 44 attributes and two classes. Class 1 has 55 and class 2 has 212 instances.
viii. Seed dataset: This is a three class problem contains 210 instances, 70 each and it has 7 attributes.
ix. Haber man's survival data: This problem has 306 instances, 3 attributes and two classes.

B. *Results and Discussion*

We have performed rigorous experiments on aforementioned bench mark problems. In order to assess our proposed modified evolutionary fuzzy clustering (MOEFC); we performed a comparative analysis with its competent methodologies that is shown in table. 1. Thus, we compared with Fuzzy C-Means clustering (FCM), Evolutionary Fuzzy C-Means clustering (EFCM) and Evolutionary Fuzzy C-Means clustering with Minkowski distance (EFCM with p). Performance measure is classification accuracy that is defined as:

$$classification\ accuracy = \frac{no.of\ correct\ matches}{total\ samples} \times 100 \quad (10)$$

In this experiment, we observed that best results are obtained at different values of parameters involved herewith. For best results, we reported the actual values of parameters. For, most of the cases, we employed the range of fuzzifier in between ($2 \leq m \leq 2.5$) and parameter p lies in between ($1 \leq p \leq 6$). €, £, ∈ are threshold values, defined by user.

We observed that for most of the problems addressed here, there is significant improvement in classification accuracy using proposed modified evolutionary fuzzy clustering as compared with variants of fuzzy clustering mentioned above.

It is clear from the table given below that modified evolutionary fuzzy clustering yields far better results than conventional fuzzy c-means clustering. However, modified evolutionary fuzzy clustering also performing better than evolutionary variants of fuzzy clustering.

TABLE.1 COMPARISON AMONG CLUSTERING VARIANTS

Dataset-Problems	**FCM (%)**	**EFCM (%)**	**EECM with p**	**MOEFC**
Breast-cancer	92.79	93.14	94.15	97.23
Fertility	63	74	76	81
Wine dataset	89.88	92.45	95.63	98.25
Blood-Transfusion	58.02	58.28	58.28	62.43
Haber man's Survival	55.228	56.53	57.18	58.22
Iris-flower	91.45	93.83	95.42	97.57
Seed dataset	90	90.47	90.89	92.54
SPECTF-heart	63.67	72.38	74.09	77.90
User-know. Modeling	52.88	53.15	54.09	56.83

The basic impetus of all clustering algorithms is the classification on N objects into C clusters and minimization of objective function. A graph has been plotted between iteration and objective-function that is shown in Figure. 2. We perceived, if numbers of iterations are increased, the objective-function decreases. In breast-cancer and in SPECTF heart data problem, there is a sudden change in objective-function. For all the remaining dataset problems it varies slightly

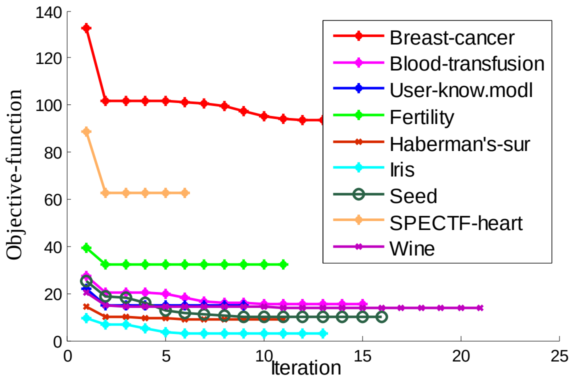

Figure. 2 Iteration Vs. objective-fun plot for all datasets

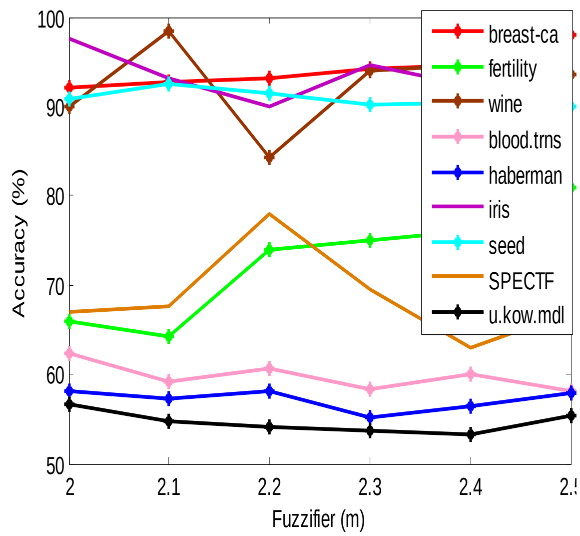

Figure. 3 Fuzzifier vs. Accuracy plot for all datasets

For each aforementioned data problem, we plotted a graph between fuzzifier and accuracy. Each problem confers its optimal result at distinct values of *m* that is incontrovertibly shown in figure. 3. Hence, we deduced the statement that suitable value of fuzzifier varies from problem to problem, i.e. for different dataset problems, fuzzifier value is different.

In figure. 4, a plot between distance parameter p and accuracy has been shown. Predominantly Minkowski parameter $p=2$ is satisfactory for most of the problems but some problems don't yield precise result at $p=2$

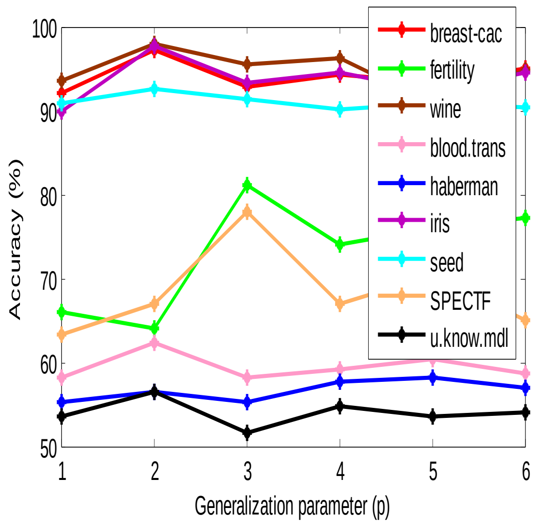

Figure. 4 P vs. Accuracy plot for all datasets

Fertility and SPECTF Heart data problem bestow ultimate results at *p=3* while Haber man's survival data problem confer optimal result at *p=5*.

In evolutionary strategy, we employed number of generations of partition matrix (*U*). For most of the problems, as we increase the no of generations of *U*,

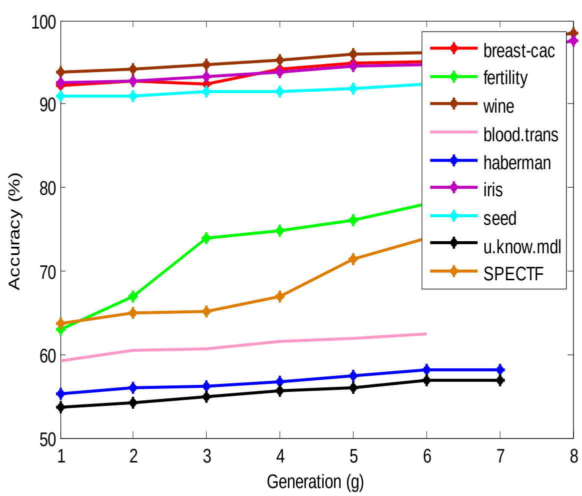

Figure. 5 Generation vs. Accuracy plot for all datasets

accuracy also increases simultaneously for certain number of generations after that it becomes constant, as shown in figure. 5. Haber man's survival data problem provides a steady result in every generation. Blood transfusion service center data problem yields an instant amend in outcome in each generation. It has been observed that different number of generations, we yield best results for different data problems. This is due to the fact that termination criteria involved in modified evolutionary fuzzy clustering contains successive refinement of fitness function as well as classification entropy that varies problem to problem. Henceforth, we obtained best outcomes of different problems on different generations.

The deployed parameter values for the best outcomes in this research work are put in a nut shell in Table. 2. Xie Beni Index constitutes a ratio of a global measure of intra-cluster variation divided by the distance between two closest clusters. It measures the validity of fuzzy clustering. Classification entropy is used to define the goodness of fuzzy clustering. Smaller value of entropy indicates less disorder in clustering, which means a better clustering.

Table.2 Actual Values of All Parameters Employed

Dataset-problem	**(m)**	**(p)**	**Itr**	**XB**	**CA**
Breast-cancer	2.5	2	14	0.0014	0.5896
Fertility	2.5	3	13	0.0016	0.6787
Wine dataset	2.1	2	22	.0042	0.8984
Blood-Transfusion	2	2	16	.0015	0.4478
Haber man's Survival	2	5	13	.0027	0.5360
Iris-flower	2	2	13	0.0111	0.6898
Seed dataset	2.1	2	16	0.0073	0.6254
SPECTF-heart	2.2	3	7	0.0019	0.6931
User-know. Modeling	2	2	8	0.00068	1.385

CONCLUSION

In this paper, we described a novel modified evolutionary Fuzzy clustering. This tactics was exploited over aforementioned bench mark problems. We perceived that class compactness is disposed by Xie Beni, as its value decreases; tends to discover a better class compactness. Classification entropy (*CA*) influences the partition matrix. We observed that minimum value of *CA* implies a good partition. Shape of the cluster is affected by the Minkowski parameter (*p*). For most of the data problems, we identify appropriate shapes at distinct values of *p.* we also employed evolutionary strategy at fuzzy partition matrix *(U)*. Accuracy increases, as we increase the number of generations and after certain number of generations it becomes constant. Finally, it has been observed that modified evolutionary fuzzy clustering that encompasses aforementioned parameters in a single system yield comparatively promising results to conventional variants of fuzzy clustering.

REFERENCES

[1] B. Simhachalam, G. Ganesan, "Possibilistic Fuzzy C-Means Clustering on Medical Diagnostic System", International conference on Contemporary Computing and Informatics, pp. 1125-1129, 2014.

[2] Z. Ye, H. Mohamadian, "Enhancing Decision Support for Pattern Classification via Fuzzy Entropy Based Fuzzy C-Means Clustering", IEEE 52nd annual conference on Decision and Control, pp. 7432-7436, 2013.

[3] J. Mathew, R. Vijaya Kumar, "Scalable Parallel Clustering Approach for Large Data Using Genetic Possibilistic Fuzzy C-Means Algorithm", IEEE International conference on Computational Intelligence and Computing research, pp. 1-7, 2014.

[4] A. G. Di Nuovo, M. Palesi, V. Catania, "Multiobjective Evolutionary Fuzzy Clustering For High Dimensional Data Problems", IEEE International conference On Fuzzy System, pp. 1-6, 2007.

[5] Bezdek, J. C., "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.

[6] J. W. Yoon, S. B. Cho, "An Efficient Genetic Algorithm with Fuzzy C-Means Clustering for Travelling Sales Man Problem", IEEE congress on Evolutionary Computation, pp. 1452-1456, 2011.

[7] Zhang Jianhua, Zeng Bo Zhang Min, Ding Lan, Dong Jun, "A Differential Evolution Optimal Fuzzy Clustering with Adaptive Adjusting Strategy", IEEE youth conference on Information Computing and Telecommunication, pp. 25-28, 2010.

[8] W. Li, G. Guo, T. Yan "Multi-objective Genetic Algorithm Based on Clustering", 2nd International conference on Intelligent System Design and Engineering, pp. 41-43, 2012.

[9] K. Deb, "Multi-objective Optimization Using Evolutionary Algorithms", Wiley, 2001.

[10] T. Van Le, "Evolutionary fuzzy clustering", In Proc, IEEE Congress on Evolutionary Computation, pp. 753-758, 1995.

[11] M. Zhang, W. Zhang, H. Sicotte, P. Yang "A New Validity Measure For A Correlation Based Fuzzy C-Means Clustering Algorithm", annual International conference of IEEE on Engineering in Medicine and Biology Socienty, pp. 3865-3868, 2009.

[12] K. M. Batainen, M. Naji, M. Saqer, "A Comparison Study between Various Fuzzy Clustering algorithms", Jordan Journal of Mechanical and Industrial Engineering, vol. 5, Issue. 4, pp. 33-343, 2011.

[13] Vivek Srivastava, Bipin K. Tripathi, Vinay K. Pathak, "An Evolutionary Fuzzy Clustering with Minkowski Distances", 18th International conference on Neural Information Processing, vol. 7063, pp. 753-760, 2011.

[14] Bo Gao, Jun Wang, "Multi-Objective Fuzzy Clustering for Synthetic Aperture Radar Imagery", IEEE Geoscience and Remote sensing Letters, Vol. 12, Issue. 11, pp. 2341-2345, 2015.

[15] X. L. Xie, G. Beni, "A Validity Measure for Fuzzy Clustering", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 13, Issue. 8, pp. 841-847, 1991.

[16] H. Zhang, A. Zhou, S. Song, X. Z. Gao, J. Zhand,"A Self Organizing Multi-objective Evolutionary Algorithm", IEEE Transaction On Evolutionary Computation, Vol. PP, Issue.99, pp. 1-1, 2016.

[17] L. Zhu, F. L. Chung, S. Wang, “Generalized Fuzzy C-Means Clustering Algorithm with Improved Fuzzy Partition”, IEEE Transaction on Systems, Man and Cybernetics, Vol. 39, Issue. 3, pp. 578-791, 2009.

[18] B. Yuan, G. J. Klir, J. F. Swan-Stone, “Evolutionary Fuzzy C-Means Clustering Algorithm”, In Proc. Int. Conference on Fuzzy Systems, pp. 2221-2226, 1995.

[19] J. C. Dunn, “A Fuzzy Relative of ISODATA Process and Its Use In Detecting Compact Well Separated Clusters”, J. Cybern, vol. 3, pp. 32-57, 1973.

[20] N. Bharill, O. P. Patel, A. Tiwari, “An Enhanced Quantum-Inspired Evolutionary Fuzzy Clustering”, IEEE symposium series on Computational Intelligence pp. 772-779, 2015.

[21] E. R. Hruschka, R. J. G. B. Compello, L. N. De Castru, “Evolutionary Search For Optimal Fuzzy C-Means Clustering”, IEEE International conference on Fuzzy Systems, vol. 2, pp. 685-690, 2004.

[22] A. G. Di Nuavo, V. Catania, “An Evolutionary Fuzzy C-Means Approach for Clustering Of Bio Informatics Data”, world congress on Computational Intelligence, pp. 2077-2082, 2008.

[23] Eduardo R. Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, Andre C. P. L. F. de Carvalho, “A Survey of Evolutionary Algorithms for clustering”, IEEE transaction on Systems, Man and Cybernetics- part c: Application and Reviews, vol. 39, Issue. 2, pp. 133-155, 2009.

[24] UCI Machine Learning Repository (www.archive.ics.uci.edu).

.

.