IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Inter-class and Intra-class Fuzzy Clustering with Pruning Algorithm

A. B. Kulkarni[1*], S.V. Bonde[2], U.V. Kulkarni [3]

[1*]CSE Department, SGGSIE&T, SRTM University, Nanded, India
[2] EXTC Department, SGGSIE&T, SRTM University, Nanded, India
[3] CSE Department, SGGSIE&T, SRTM University, Nanded, India

Corresponding Author: kkkarun@yahoo.com, Tel.: +91-98334-28466

Available online at: www.ijcseonline.org

Accepted: 22/May/2018, Published: 31/May/2018

Abstract— The paper proposes a new supervised fuzzy clustering algorithm based on inter-class and intra-class clustering technique to create the clusters i.e. fuzzy hyperspheres (FHSs) and pruning technique to prune redundant FHSs which are camouflaged by the other FHSs of the same class. The proposed clustering technique finds the centroid and the width of the FHS based on the spread of inter-class patterns and then groups intra-class patterns using fuzzy membership function, whereas the pruning technique creates the optimal number of FHSs from the FHSs created in the earlier stage. This algorithm is independent of parameters, limits the interference of outliers and converges quickly to create an optimal number of clusters. The main feature of the proposed fuzzy clustering algorithm is that it camouflages the clustered patterns giving 100% accuracy for any training dataset. The performance of the proposed algorithm is tested on eleven benchmark datasets and it is observed that the proposed algorithm results are superior and comparable with classifiers using clustering algorithm.

Keywords—Fuzzy clustering, Fuzzy membership function, Fuzzy hyperspheres, pruning.

I. Introduction

Clustering is a task whose goal is to determine a finite set of categories (clusters) to describe a data set according to similarities among its objects. So clustering is ubiquitous, also called as exploratory data analysis, which works with labelled or unlabelled data to form clusters [1].

The applicability of clustering is manifold and immensely increasing in pattern recognition [2], image segmentation [3], text document analysis [4], speech processing [5], medical diagnosis [6], Content-based Image Retrieval [7], wireless sensor network [8] etc. The importance of clustering has grown fast and persistently over the past recent years in engineering and scientific applications. There are considerable publications devoted to clustering analysis over the past decade. Different Researchers have proposed and used various approaches to solve the scientific applications in the real world. Almost all clustering algorithms have the flaws and are suitable for certain data types. So there is a continuous demand for researching different kinds of clustering algorithms. This paper proposes an algorithm which is independent of almost all the data type and can be used in real-world application.

The rest of the paper is organized as follows. The following section II describes the related work i.e. various clustering algorithms and the limitations in the recent clustering algorithm in brief. In section III maximum spread fuzzy clustering algorithm with pruning to construct FHSs is described in detail. Results of three case studies are discussed in section IV. Finally, section V concludes the paper with future scope.

II. Related Work

Clustering algorithms are commonly accepted as optimal quantization approaches but they are very time-consuming. Moreover, the clustering algorithms suffer from their dependence on initial conditions. In most of the applications, one specific initial condition is chosen to present the results. However, using other initial conditions can change the performance of the algorithm dramatically. So different starting points and criterion usually lead to different taxonomies of clustering algorithms [1]. The other important aspect is the standards we should use to determine the closeness or how to measure the distance between the objects, an object and a cluster, or a pair of clusters. A rough but widely agreed frame is to classify clustering techniques as hierarchical clustering [9] and partitioned clustering [10]. Partitioning methods can be divided into hard clustering and fuzzy clustering. Hard clustering provides a hard partition in which each object in the dataset is assigned to one and only one cluster [11]. For Fuzzy clustering, the restriction is relaxed, and the object can belong to all of the clusters with a certain degree of membership [12]. Fuzzy min-max neural network by Simpson laid to some dimension in formation of clusters [13]. Based on this fuzzy hyperline segment clustering neural network was developed as the improvement in clustering Fisher Iris data [14]. Neural network based clustering has been dominated by Self organizing fuzzy maps (SOFMs) and adaptive resonance theory (ART) [15]. Fuzzy ART (FA) benefits the incorporation of fuzzy set theory and ART [16]. Many clustering algorithms have been developed in the past sixty years, among these algorithms, the k -means algorithm is one of the oldest and most commonly used clustering algorithms [17]. Cluster validation is another aspect for various algorithms and applications [18].

The proposed work compares and overcomes the limitations in the clustering algorithms used in radial basis function neural network proposed in [20] [21] like adjusting the static parameters, training accuracy, and number of clusters.

III. METHODOLOGY

The Maximum spread fuzzy clustering with pruning (MSFCP) algorithm to construct the optimum number of FHSs is described in the following subsection. It basically consists of two steps: a) Creation of FHSs [19] b) Pruning of FHSs.

a) Creation of FHSs: Let Z be the training set containing P training pairs (X_h, d_h), where X_h is the h^{th} input pattern and d_h represents the desired output for X_h.

Consider α_k be the patterns of class C_k which is the subset of set Z, then following steps are executed for K classes and k is one of the class varying from $k = 1, ..., K$.

Step 1: The distance between the patterns of k^{th} class with inter-class patterns is determined and stored in A^k .

$$A^k = \left[\left\| X_i - X_j \right\|\right]_{\alpha_k \times t_k}, i = 1, 2, ..., \alpha_k \text{ and } j = 1, 2, ..., t_k, \quad (1)$$

$$\text{where } X_i \in C_k,\ X_j \notin C_k, \text{and } t_k = P - \alpha_k.$$

Step 2: The minimum of the distance of each pattern of k^{th} class with the inter-class pattern is determined as

$$B^k = \min\left(A^k\right) \quad (2)$$

Step 3: Using B^k , the pattern X_j^k having maximum spread is considered to be the centroid of FHS with an initial radius equal to

$$g^k = \max\left(B^k\right) \quad (3)$$

Step 4: Using X_j^k, g^k and the membership function the intra class patterns are clustered .The membership function for the FHS is defined as

$$m_j\left(X_h, C_j, r_j\right) = f(l, r_j) \quad (4)$$

where $X_h = (x_{h1}, x_{h2}, \ldots, x_{hn})$ is an input pattern, r_j is a radius of j^{th} FHS H_j with a centroid $C_j = [c_{j1}, c_{j2}, ..., c_{jn}]$. $f()$ is defined as:

$$f(l, r_j) = \begin{cases} 1 & l \le r_j \\ r_j / l & otherwise \end{cases} \quad (5)$$

where l is an Euclidean distance between X_h and C_j .

Step 5: Determine the final radius of this FHS by following equation

$$r_j^k = \begin{cases} \max\limits_{i=1}^{n_j} d_i, & for\, n_j > 1 \\ \dfrac{g^k}{2}, & for\, n_j = 1 \end{cases} \quad (6)$$

Where d_i is the distance between X_i^k pattern and the centroid X_j^k of newly created FHS, n_j is the total number of patterns clustered by the FHS.

Step 6: Calculate $\alpha_k \leftarrow \alpha_k - n_j$.

Step 7: If $\alpha_k \neq 0$ delete the corresponding rows of the clustered patterns in A^k and go to step 2, else go to step 8.

Step 8: Repeat the above steps for all classes i.e. till $k \neq K$.

b) Pruning of FHSs: Let $Q_1, Q_2, ..., Q_K$ be the set of FHSs clustering more than one pattern for each class, while $S_1, S_2, ..., S_K$ be the respective set of FHSs which clusters only one pattern.These sets are created using step (a) of MSFCP algorithm. The set Q_k and S_k represent the FHSs of class k and are defined as $Q_k = \left[q_{k1}, q_{k2}, ..., q_{kn}\right]$ where q_{ki} represents the i^{th} FHS of class k and $i = 1, 2, ..., n$ and on the same basis we define $S_k = \left[s_{k1}, s_{k2}, ..., s_{km}\right]$.

The following steps are carried out to prun the FHSs in S_k for all classes where $k = 1, 2, ..., K$.

Step 1:In this step the FHSs in S_k are validated by using the membership function associated with the FHSs in Q_k. The FHSs in S_k are pruned/eliminated if any of the FHSs in Q_k gives the highest membership value in comparision with Q_j and S_j where $j \neq k$.

Step 2: The final FHSs for the class k is given by the relation $Q_k = [Q_k\ S_k^{'}]$. where $S_k^{'}$ represents the set of unpruned FHSs from S_k .

Step 3: Repeat the above steps till $k \neq K$.

Testing of MSFCP algorithm:

Once the final clusters i.e. FHSs are created using the MSFCP algorithm, then the performance in terms of recognition rate is tested using the following procedure.

1) Apply the input pattern from the data set to all the created FHSs.

2) Using membership function, calculate the membership value of the input pattern for each FHS.

3) The input pattern is said to belong to the class whose FHS gives the maximum membership value.

IV. RESULTS AND DISCUSSION

The MSFCP algorithm has been implemented in Matlab 2016a. To evaluate its performance three case studies along with obtained results are discussed in the following sub-sections.

Case Study 1:

To have better understanding of MSFCP algorithm for creating precise and optimal number of FHSs, a 2-dimensional 3 class example is illustrated. The training set consists of twenty four 2D patterns: (1, 5), (2, 4.5), (0.75, 6), (1, 7.5), (1.5, 7), (0.75, 8), (1.5, 8.5), (1, 9) belonging to class 1, (1, 1), (2, 1.5), (3, 1.5), (0.75, 3), (1.5, 2.5), (2, 3.5), (3.25, 4), (4, 1) belonging to class 2 and (3.7, 6), (4, 5), (4, 7), (3, 8), (3.75, 8), (3.5, 8.5), (3.75, 9), (3, 9) belonging to class 3. The scatter plot of these patterns is shown in Fig. 1.

By using step (a) the MSFCP algorithm constructed four FSHs; two FSHs for class 1 and one FSH for class 2 and class 3 each. The centroids of class 1 FSHs are (0.75, 6.0), and (1.0, 9.0) with radii 2.6101, and 1, respectively. The class 2 FHS centroid is (4, 1) with radius 3.8161. Similarly, class 3 FSH centroid is (4.0, 7.0) with radius 2.2361. The Fig. 6 shows the constructed FSHs for all 3 classes.The detail description of these created FSHs is explained below.

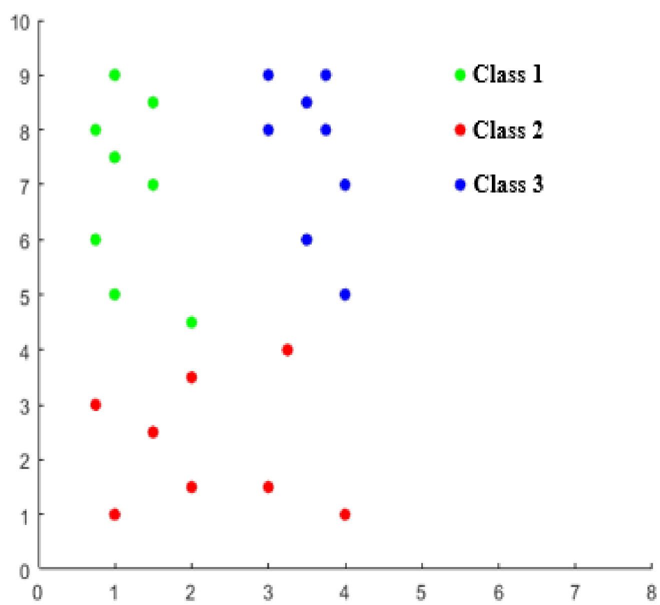

Figure 1: 2-D Scatter plot of patterns

Initially as per the MSFCP algorithm, the clustering process for the patterns in class 1 is done. As per the step 1, the inter class distance of class 1 patterns with class 2 and class 3 is calculated. Using step 2 and step 3 the pattern (.75, 6) of class 1 is considered as centroid as it has maximum spread with initial radius 2.7951 due to the pattern (4, 6) of class3. Later by step 4, (.75, 6) as the centroid of first FHS with the radius equal to 2.7951, cluster the patterns of class 1 using the defined membership function. The first FHS clusters all the patterns except (1, 9) as shown in Fig. 2. Then using step 5 we calculate the final radius i.e. 2.6101 equal to the maximum distance between the centroid and clustered pattern (1.5, 8.5) which is shown in Fig. 2.

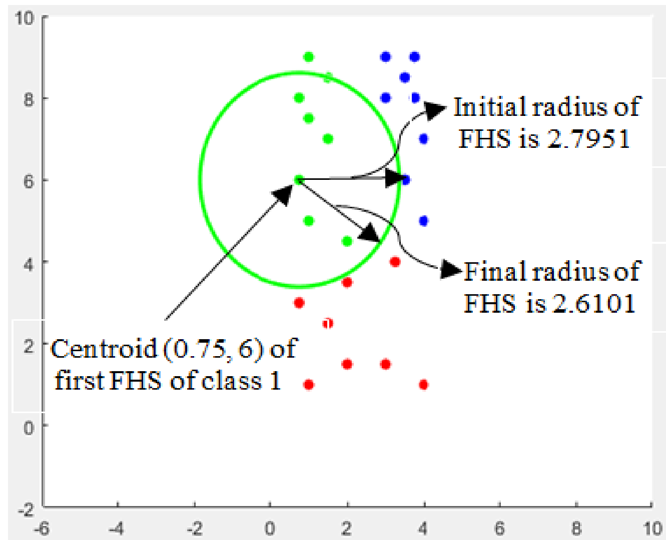

Figure 2: First FHS of class 1 with centroid and radius

Since one more pattern of class 1 is not clustered so as per step 7 the same process is repeated and the second FHS for this class is created as shown in Fig. 3. This FHS clusters only one pattern, so as per step 5 the final radius assigned is equal to half of the initial radius. As per step 7 the process of creation of FSHs for class 1 is over.

.

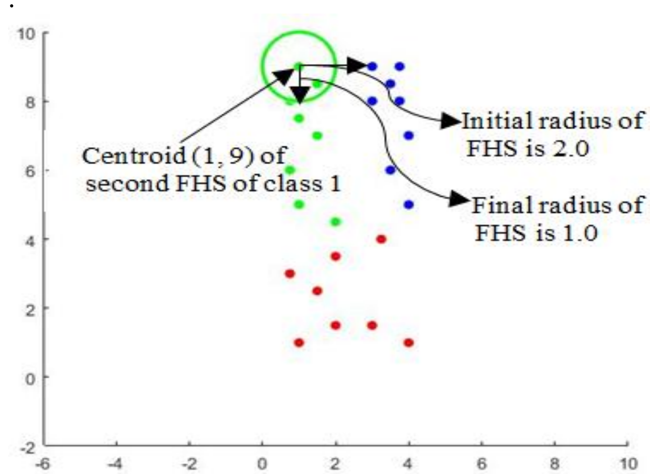

Figure 3: Second FHS of class 1 with centroid and radius

As per step 8 the same procedure is repeated for other classes. So for class 2 as per the MSFCP algorithm, the pattern (4, 1) is selected as centroid with initial radius to be 4.0, due to the pattern (2, 4.5) of class 1. Then the final radius 3.8161 is adjusted by using step 5 because of the maximum distance between the centroid and the pattern (.75, 3) of class 2 is 3.8161 which is shown in Fig. 4.

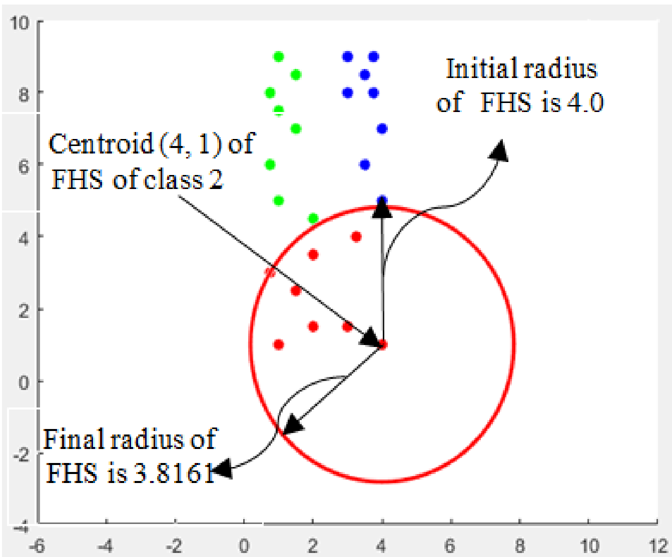

Figure 4: FHS of class 2 with centroid and radius

As all the patterns of class 2 are clustered in one FSH we proceed for class 3. For this class, FHS with centroid (4, 7) and final radius as 2.2361 is created due to the pattern (3.5, 8.5) of class 3 as shown in Fig. 5.

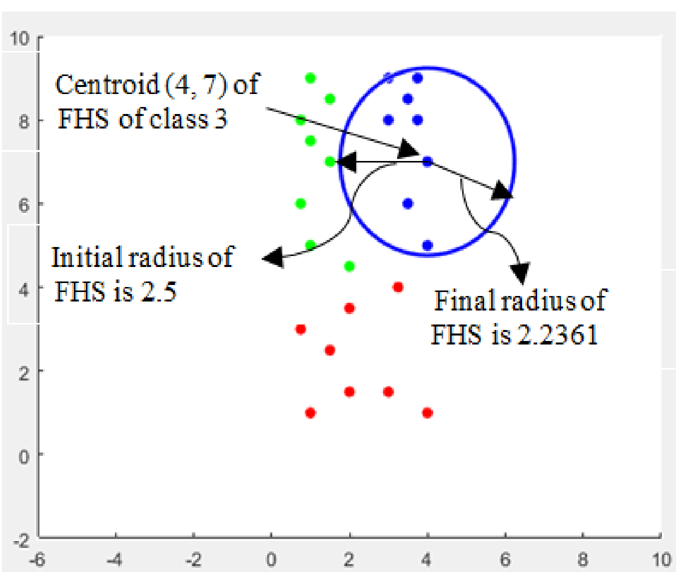

Figure 5: FHS of class 3 with centroid and radius

Thus, we have created four FHSs for all three classes i.e. two FHSs for class 1, one for class 2 and class 3 each.The Fig. 6 shows these FHSs for all three classes.

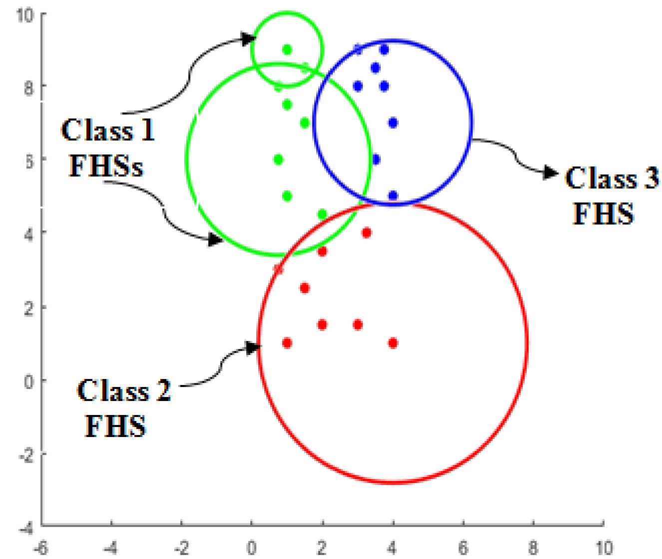

Figure 6: FHSs for all 3 classes without pruning

Now using step (b) of MSFCP algorithm, the second FHS of class 1 is pruned as the first FHS of class 1 gives maximum membership value in comparision with the other FHS of class 2 and class 3. Pruning is not required for class 2 and class 3, since these classes don't have the FHSs having clustered single pattern. The constructed FHSs for 2-D example of 3 classes using MSFCP algorithm is shown in Fig. 7

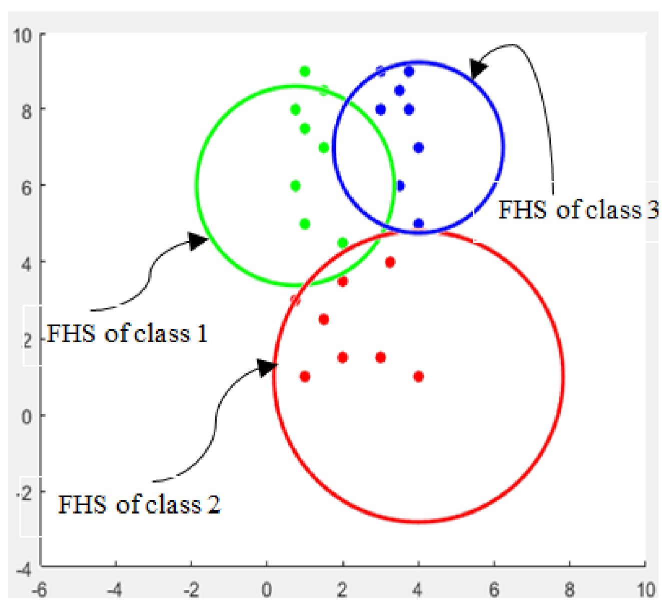

Figure 7: Final FHSs for all three classes after pruning

Case Study 2:
The performance of proposed algorithms is verified using ten UCI datasets. The experimental procedure in [20] is followed, to have a fair comparison between the MSFCP algorithm and other classifiers. The average percentage 5-fold validation test accuracies are tabulated in Table 1 along with the results given in [20]. The results show that MSFCP algorithm is superior for five datasets and comparable with remaining datasets.

Table 1: Average percentage 5-fold validation test accuracies

Dataset	MSFCP	RBF	RBF-R	RBF-N	RBF-WTA
Hepatitis	**88.2**	65.0	81.9	81.1	82.1
Zoo	92.4	83.8	95.2	94.3	96.2
Glass	**75.0**	38.7	66.1	66.3	69.1
Heart	77.0	73.5	81.9	80.5	80.6
Ecoli	**83.5**	69.5	78.5	79.3	81.0
Liver	**69.6**	53.8	62.2	62.8	61.0
Ionosphere	90.0	81.5	95.5	95.2	94.3
Monks-3	83.4	97.5	99.0	95.8	68.6
Breast	96.2	94.1	96.3	96.4	97.0
Pima	**76.9**	71.0	75.3	72.1	73.8

Case Study 3:
As we are aware, the main issue with the clustering algorithms is the number of clusters formed. With this perspective, the performance of MSFCP algorithm is compared with respect to the average number of clusters and the corresponding recognition error rate for the respective dataset. Table 2 and 3 show the comparison with the other classifiers as specified in [21] with respect to recognition error and the average number of clusters. Table 2 and 3 shows the results which are comparable, and though the number of clusters/FHSs are more for some dataset, still the computation time will be less as compared to other classifiers. The results for number of FHSs created for Ionosphere data set is very less in comparision with other classifiers.
The MSFCP algorithm shows good data visualization in the clustering process and guarantees the convergence of the algorithm after few iterations. Adjustment of static parameter is not required during training and testing. The proposed algorithm gives 100% efficiency for all above datasets.

Table 2: Comparision of Recognition error rate

Dataset	Recognition error rate					
	MSF CP	**Complex -valued**	**Real valued**	**RBF -R**	**RBF -N**	**RBF-WTA**
Thyroid	7.9	2.95	3.78	4.4	5.3	3.7
Heart	23.0	17.08	19.68	18.1	19.5	19.4
Ionosphere	10.0	3.7	7.7	7.14	6.13	4.5
Breast	3.8	2.9	1.1	3.7	3.6	3.0

Table 3: Comparision of average number of clusters/FHSs

Dataset	Average number of clusters/FHSs					
	MSF CP	**Complex -valued**	**Real valued**	**RBF -R**	**RBF -N**	**RBF-WTA**
Thyroid	15	19.65	20.78	15.1	18.7	14.6
Heart	53.2	44.12	46.15	24	27	46
Ionosphere	37.8	117.12	116.45	65	48	66.6
Breast	37.4	28.12	29.48	40	35	40

V. Conclusion and Future Scope

The proposed MSFCP clustering algorithm creates fuzzy hyperspheres i.e. clusters on the basis of inter-class and intra-class fuzzy membership metric with the maximum spread in two steps. The created FHSs are characterized by the membership function, assuring 100% training efficiency for any dataset. During the first step, the MSFCP algorithm creates the possible number of FHSs and then by second step it performs the pruning operation to reduce the number of FHSs. The pruning operation removes the FHSs clustering the single pattern with the help of the FHSs of the same class without affecting training efficiency. Randomizing the order of clustering may improve the test efficiency as well as the number of FSHs formed by MSFCP algorithm. If the outliers handling techniques can be implemented before clustering then it may increase the test efficiency and reduce the overall number of FSHs.

References

[1] K. Rose, F. Guerewitz, G. Fox, "*A Deterministic annealing Approach to Clustering*", Pattern Recognition Let., Vol. 11, Issue. 9, pp. 589-594, 1990.

[2] L. Bai, J. Liang, C. Dang, F. Cao, "*A Novel Fuzzy Clustering Algorithm With Between-Cluster Information for Categorical Data*", Fuzzy Sets Syst., Vol. 215, pp. 55-73, 2013.

[3] F. Tung, A. Wong, D. A. Clausi, "*Enabling Scalable Spectral Clustering for Image Segmentation*", Pattern Recogn., Vol 43, Issue. 12, pp. 4069-4076, 2013.

[4] Y. Yan, L. Chen, W. C. Tjhi, "*Fuzzy Semi-Supervised Co-Clustering for Text Documents*", Fuzzy Sets Syst., Vol 215, pp. 74-89, 2013.

[5] B. Sun, W. Liu, Q. Zhong, "*Hierarchical Speaker Identification Using Speaker Clustering*," Int. Conf. on Natural Language Processing and Knowledge Engineering, pp. 299-304 2003.

[6] B. Dogan, M. Korurek, "*A New Ecg Beat Clustering Method Based On Kernelized Fuzzy C-Means And Hybrid Ant Colony Optimization for Continuous Domains*", Appl. Soft Comput.,Vol 12 , Issue. 11, pp. 3442–3451, 2012.

[7] Y. Chen, J. Wang, And R. Krovetz, "*Clue: Cluster-Based Retrieval Of Images By Unsupervised Learning*", IEEE Trans. on Image Processing, Vol.14, Issue. 8, pp. 1187–1201, 2005.

[8] C. R. Lin And M. Gerla, "*Adaptive Clustering for Mobile Wireless Networks*", Journal on Selected Areas in Communication, Vol. 15, Issue. 7, pp.1265-1275, 1997.

[9] R.N. Dave, R. Krishnpuram, "*Robust Clustering Method: A Unified View*", IEEE Trans. Fuzzy System, Vol. 5, Issue. 2, pp. 270-293, 1997.

[10] J. C. Bezdek, "*Pattern Recognition With Fuzzy Objective Function Algorithms*", Plenum press, New York, 1981.

[11] L. Kaufman, P.J. Rousseeuw, "*Finding Groups In Data: An Introduction to Cluster Analysis*", Wiley, Hoboken, 2005.

[12] N. R. Pal, K. Pal, J. M. Keller, J. C. Bezdek, "*A Possibilistic Fuzzy C-Means Clustering Algorithm*", IEEE Trans. Fuzz,Y Syst., Vol.13, No.4, pp. 508-516, 2005.

[13] Simpson P. K.,"*Fuzzy Min-Max Neural Networks Part-2: Clustering*", IEEE Trans. Fuzzy System, Vol. 1, Issue. 1, pp.32-45, 1993.

[14] U. V. Kulkarni, T. R. Sontakke, A. B. Kulkarni, "Fuzzy Hyperline Segment Clustering Neural Network", Electronics Letters, Vol.37, Issue. 5, pp. 301-303, 2001.

[15] J. C. Bezdek, N. R. Pal, "*Generalized Clustering Networks And Kohonen's Self-Organizing Scheme*", IEEE Neural Networks, Vol. 4, Issue. 4, pp. 549-557, 1993.

[16] G. Carpenter, S. Grossberg, N. Maukuzon, J. Reynolds, And D. B. Rosen, "*Fuzzy Artmap: A Neural Network Architecture for Incremental Supervised Learning Of Analog Multidimensional Maps*", IEEE Trans. Neural Networks,Vol. 3, Issue. 5, pp. 698-713, 1992.

[17] A. Likas, N. Vlassis, Verbeek, "*The Global K-Means Clustering Algorithm*", Pattern Recog. Let., Vol. 36, pp. 451-461, 2003.

[18] D. W. Kim, K. H. Lee, D. Lee, "*Fuzzy Cluster Validation Index Based On Inter-Cluster Proximity*," Pattern Recognition Letters, Vol. 24, Issue. 15, pp. 2561-2574, 2003.

[19] A. B. Kulkarni, S. V. Bonde, U. V. Kulkarni, "*A Novel Fuzzy Clustering Algorithm for Radial Basis Function Neural Network*", International Journal on Future Revolution in Computer Science and Communication Engineering, Vol. 4, Issue. 4, pp.751-756, 2018.

[20] M. Rouhani, D. S. Javan, "*Two Fast And Accurate Heuristic Rbf Learning Rules for Data Classification*", Neural Networks, Vol.75, pp. 150-161, 2016.

[21] Yuanshan Liu, He Huang, Ting Wen Huang B, Xusheng Qian,"*An Improved Maximum Spread Algorithm With Application to Complex-Valued Rbf Neural Networks*", Neurocomputing, Vol. 216, pp. 261-267, 2016.

Authors Profile

A. B. kulkarni has obtained Bachelor of Engineering degree in Electronics from Marathwada University, Aurangabad, Maharashtra, India in 1990. He completed Master of Engineering in Electronics (Specialization in computer technology) from Swami Ramanand Teerth Marathwada University Nanded, Maharashtra, India in 2001. Currently he is the research scholar in Computer Science and Engineering Department at SGGSIE&T (Autonomous), Nanded, Maharashtra, India. He has published 15 research papers in reputed National and International Journals. His areas of interest include image processing, soft computing, clustering, and machine learning. He has nearly 28 years of teaching experience.

Dr. S. V. Bonde has obtained Bachelor of Engineering degree in Electronics from Marathwada University, Aurangabad, Maharashtra, India in 1988. He completed Master of Engineering in Electronics in 1994 from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India. He completed his Ph.D. in Biomedical Engineering from Indian Institute of Technology Bombay, Mumbai, Maharashtra, India in 2004. He is currently working as Professor in Electronics and Telecommunication Engineering Department at SGGSIE&T (Autonomous), Nanded, Maharashtra, India. He has received Sushrutha Award for best paper in Doctoral category of International Conference on Biomedical Engineering (BIOVISION 2001) at Indian Institute of Science, Bangalore and Biomedical Society of India. He has published many research papers in reputed National and International Journals. His areas of interest include image processing, wavelet transform and signal processing. He has more than 29 years of teaching experience.

Dr. U. V. Kulkarni has obtained Bachelor of Engineering degree in Electronics from Marathwada University, Aurangabad, Maharashtra, India in 1987. He completed Master of Engineering in system software from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India in 1992. He has completed his Ph.D. in Electronics and Computer Science Engineering in 2002 from Swami Ramanand Teerth Marathwada University Nanded, Maharashtra, India. He is currently working as Professor and Head in Computer Science and Engineering Department at SGGSIE&T (Autonomous), Nanded, Maharashtra, India. He has received National Level Gold Medal and Computer Engineering Division Prize for the paper published in the Journal of Institution of Engineers, titled as Fuzzy Hypersphere Neural Network Classifier, in May 2004 and the best paper award for the research paper presented in international conference held at Imperial College London, U.K., 2014. He has published many research papers in reputed National and International Journals. His areas of interest include microprocessors, data Structures, distributed systems, fuzzy neural networks, and pattern classification. He has more than 30 years of teaching experience.

