Research Journal of Applied Sciences 8 (9): 425-434, 2013
ISSN: 1815-932X

Improved Fuzzy Ant-Based Clustering: A Nonparametric Balance Between Exploitation and Exploration

Phichete Julrode and Siriporn Supratid
Faculty of Information Technology, Rangsit University, 12000 Pathum-Thani, Thailand

Abstract: Fuzzy ant-based clustering algorithm has been efficiently employed to serve real-world applications. Although, ant-based clustering algorithm can relieve the fast convergence during the search, limitation of such an algorithm to overcome the problems of local optimal traps along with divergence of the search are still non-trivial. Striking the balance between exploitation and exploration of the search is one of the significant keys to overcome such problems thus leads to achieve the global optimal solution. Nevertheless, arbitrarily defined parameters are usually used to control the cycle of exploitation and exploration mechanisms thus may lead to a biased and overly optimistic learning process. This study proposes an improved version of the fuzzy ant-based clustering. The objective is to apply a nonparametric method of balancing exploitation and exploration search during ant-based clustering, aiming to accomplish the global optimal solution. The criteria of performance evaluation rely on F-measures, FCM objective degree, Xie-Beni validity index and runtime as well. The experimental results, based on both real-world and artificial data sets indicate the high performance of the proposed method over the comparatively effective clustering algorithms.

Key words: Fuzzy c-means, ant-based clustering, nonparametric, exploration, exploitation, divide and conquer

INTRODUCTION

Clustering is one of the most important unsupervised learning techniques (Rojas, 1996; Herrero *et al*., 2011). It organizes a set of sample cases into similar groups called clusters. The objects within one cluster are highly similar and dissimilar with the objects in other clusters. Clustering is widely applied in several application fields such as pattern recognition (Webb, 2002), data mining (Tan *et al*., 2004), machine learning (Alpaydm, 2004), etc. For solving clustering problems, efficient approaches such as Self-Organizing feature Maps (SOM) (Kohonen, 1995), Average Linkages (AL) (Hastie *et al*., 2009) have been successfully applied. On the other side, k-means (MacQueen, 1967) a partitional type of clustering employs simply basic idea relating to find cluster centers then refining them. The soft clustering version of k-means, Fuzzy C-Means (FCM) (Dunn, 1973; Bezdek, 1981) allows each sample cases belonging to two or more clusters with different degrees of membership thus FCM is well applied to real-world applications. Nevertheless, FCM is sensitive to initialization and can be easily trapped into local optimal solutions. In order to relieve such a difficulty, most of the researches have been proposed, aiming to integration between soft clustering and powerful evolutionary optimization algorithms, i.e., FCM and particle swarm optimization (Gan *et al*., 2009; Izakian and Abraham, 2011) as well as differential evolution and k-Harmonic means which relies on harmonic means (Kao *et al*., 2008; Tian *et al*., 2009; Gong *et al*., 2009; Supratid and Julrode, 2011). Ant-based algorithm has been developed using swarm intelligence principles that emphasize distributiveness, direct or indirect interactions among relatively simple agents, flexibility and robustness (Bonabeau *et al*., 1999). By such competent characteristics, ant-based clustering more relieves the fast convergence during searching process than several other evolutionary approaches. Fuzzy ant-based clustering was primarily proposed by Kanade and Hall (2003). The ants search for optimal set of clusters using 2D grid. In order to accomplish the search, the ants move the similar sample case items into the same cluster and those dissimilar into the different ones. Then, the cluster centers, found by the ants are refined using FCM. In later version (Kanade and Hall, 2007), the ants perform clustering tasks on a basis of cluster centroids position. The ants move the cluster centers, not the sample case items to relocate the cluster centroids in the feature space. A particular partition, consisted of optimal set of clusters is discovered. The latter algorithm called fuzzy ant-based clustering with cluster centroids positioning has fewer number of controlling parameters than the previous version where various thresholds to merge and segregate the sample cases on 2D grid are to be used. Like the 2D grid version,

Corresponding Author: Siriporn Supratid, Faculty of Information Technology, Rangsit University, 12000 Pathum-Thani, Thailand

FCM is subsequently applied next to the ant clustering in order to achieve better cluster results. However, the globally best solution cannot be attained if most of the ants exploit the search space around the optimally best solution. On the other side, if the ants explore the search space to get better solutions by enhancing the diversity of solutions, the clustering may need more time to converge. Therefore, it is necessary to strike the balance between exploration and exploitation for achieving globally best solution (Bonham, 1999; Chang, 2004; Fang *et al*., 2009). Several researches attempt to accomplish such an equilibration based on some given parameters. This may lead to a biased and overly optimistic learning process; thus limit the usefulness of the model (Kiang, 2003). In a nonparametric algorithm, none of arbitrarily setting parameters is used to control or direct the algorithm functions. The learning is automatically adjusted by the algorithm itself. The nonparametric learning algorithm, proposed by Li and Yeh (2008) is employed to speed up stabilizing the learning task and can dynamically improve deriving knowledge. Another nonparametric algorithm combines learning technique and a linear programming approach (Pai *et al*., 2012) this combination considerably improves the classification accuracy as well as reliability. However, nonparametric balance between exploration and exploitation has not been much investigated.

This study proposes an improved version of Fuzzy Ant-Based Clustering algorithm. The objective is to apply a nonparametric method of balancing exploitation and exploration search during ant-based clustering, aiming to accomplish the global optimal solution. None of arbitrarily setting parameters is used to control the mechanisms of exploitation and exploration. Here, the exploration and exploitation complies the regulation of divide and conquer principle (Rugina and Rinard, 2001). The criteria of performance evaluation rely on F-Measures, FCM objective degree and Xie-Beni validity index (XB). Additionally, runtimes of the algorithms are provided. The experiments are taken on six benchmarks real-world and two artificial data sets. The comparison tests are performed on the proposed method, IFAC against the former fuzzy ant-based clustering with cluster centroids positioning, ant-based clustering alone as well as some other types of effective clustering algorithms such as SOM and AL.

FUZZY ANT-BASED CLUSTERING WITH CLUSTER CENTROIDS POSITIONING (FAC)

The fuzzy ant-based clustering with cluster centroids positioning (FAC) was originally proposed by Kanade and Hall (2003). It is a combination between Ant-based clustering (ANT) and FCM aiming to search for optimal partition of cluster centers. Initially, the feature values are normalized between 0 and 1. An ant is assigned to a particular feature of a cluster center in a partition. To search for the new partition of clusters, ants randomly move the clusters in a corporative manner. Two directions are defined for the random movement of the ant. The positive direction is when the ant is moving in the feature space from 0 to 1 and the negative direction is when the ant is moving in the feature space from 1 to 0. If during the random movement the ant reaches the end of the feature space, the ant reverses the direction. After moving the cluster centers for a fixed number of iterations, the quality of the partition is evaluated using FCM objective function specified in Eq. 1:

$$\text{FCM\_ObjectiveFunction} = \sum_{i=1}^{N}\sum_{k=1}^{K}\mu_{ik}\left\|x_i - c^k\right\|^2 \quad (1)$$

where, μ_{ik} represent membership of x_i which is sample i in cluster c^k. For crisp data, μ_{ik} is zero if x_i is in cluster c^k and is one if not. After the ant process is terminated, the best partition achieved is submitted to FCM to carrying on the clustering and attain the better result.

IMPROVED FUZZY ANT-BASED CLUSTERING (IFAC)

It is still highly possible for either ANT or FAC that ants may exploit the search space only around the optimal best solution or they may overly explore diverse solutions. Additionally, it is hard to find the appropriate parameters for controlling the balance between those exploitation and exploration mechanisms. Thus, globally best solution may not be achieved. Therefore, an improved version of fuzzy ant-based clustering is proposed here. The aim is to overcome local optimal problems as well as the divergence of solutions using a technique of nonparametric balance between exploitation and exploration. The functions of exploration and exploitation here follow the effective divide-and-conquer principle (Rugina and Rinard, 2001). The algorithm of the IFAC is described in Fig. 1.

According to Fig. 1, data are normalized at the initial step. The two initial partitions, P_1 and P_2 are randomly selected and respectively represented by the matrices: $[f_d(c_{P1}{}^k)]_{D\times K}$ and $[f_d(c_{P2}{}^k)]_{D\times K}$. Such matrices are composed of feature d in cluster c^k, d = 1, ..., D and k = 1,..., K where D and K are the number of features and the number of clusters consecutively. The principle of divide and conquer is implemented in the iteration of IFAC. The

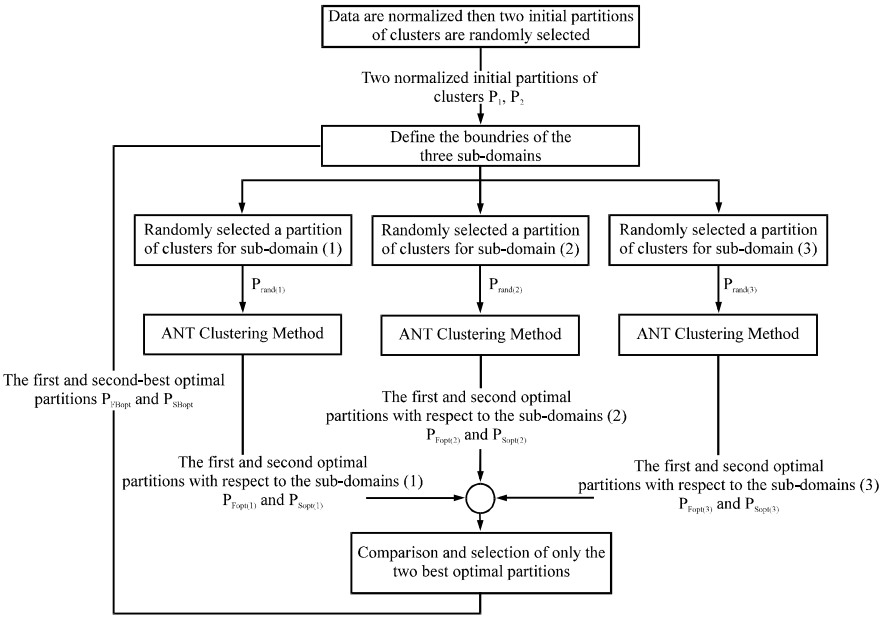


Fig. 1: The overall process of the Improved Fuzzy Ant-based Clustering algorithm (IFAC)

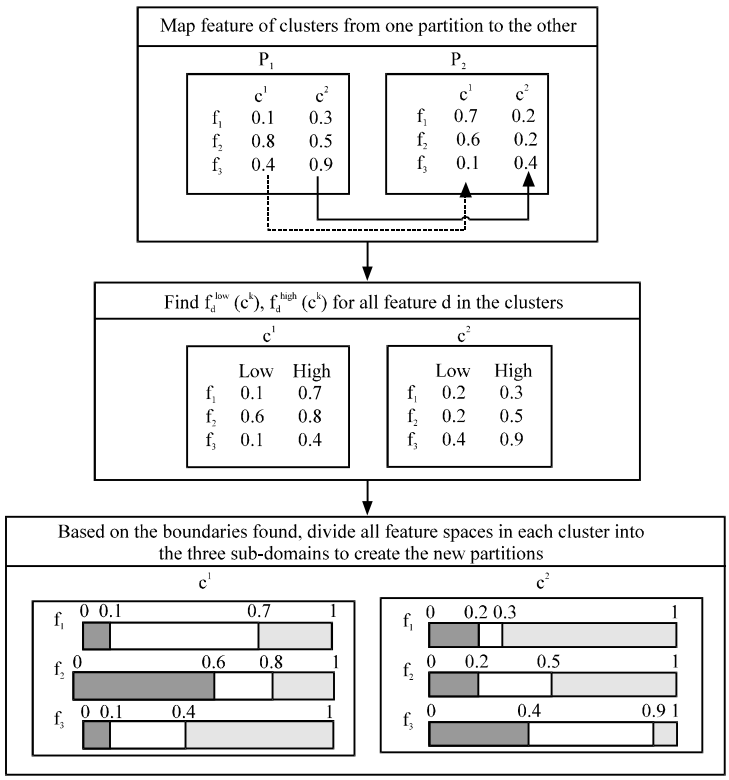


Fig. 2: An example of finding the boundaries $f_d^{low}(c^k)$ and $f_d^{high}(c^k)$ of the three sub-domains

domain of an individual feature space d, belonging to cluster c^k in each partition P_1 and P_2 is divided into three sub-domains, defined by $f_{d(1)}(c^k)$, $f_{d(2)}(c^k)$ and $f_{d(3)}(c^k)$.

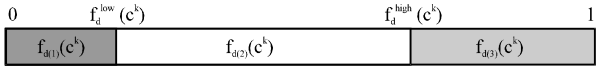
Each feature space $f_d(c^k)$ in each partition is divided into three sub-domains. $f_d^{low}(c^k)$ and $f_d^{high}(c^k)$ and refer to boundaries, low and high associated with those

sub-domains. Figure 2 delineates an example of finding such boundaries and three sub-domains. Partition P_1 and P_2 are comprised of two clusters c^1 and c^2 each of those consist of three features f_1, f_2 and f_3. The features in the corresponding clusters are mapped from one partition to the other. The smaller value of feature d with respect to the mapped clusters is specified as $f_d^{low}(c^k)$ and the larger one is designated $f_d^{high}(c^k)$. This is supported by Eq. 2 and 3:

$$f_d^{low}(c^k) = \min(f_d(c_{P_1}^k), f_d(c_{P_2}^k)) \quad (2)$$

$$f_d^{high}(c^k) = \max(f_d(c_{P_1}^k), f_d(c_{P_2}^k)) \quad (3)$$

Based on the boundaries found, an individual feature space in the cluster is divided into three sub-domains as seen at the bottom of Fig. 2. Two gray areas of sub-domains, respectively refer to $f_{d(1)}(c^k)$ and $f_{d(3)}(c^k)$, relevant to each feature space $f_d(c^k)$:



The white area of sub-domain relates to $f_{d(2)}(c^k)$. Such $f_{d(2)}(c^k)$ covers the space in between the corresponding clusters in the two best optimal partitions. Hence, such sub-domain $f_{d(2)}(c^k)$ can be regarded as an exploitation space. Whereas, the other two sub- domains $f_{d(1)}(c^k)$ and $f_{d(3)}(c^k)$ are located out of scope of the best optimal partitions. Consequently, such two latter sub-domains are counted as exploration spaces. This is a noticeable evidence of exploration and exploitation mechanisms, performed in IFAC. Afterwards, three randomly selected partitions, $P_{rand(1)}$, $P_{rand(2)}$ and $P_{rand(3)}$ are created according to a specific range of possible feature values, indicated at the bottom of Fig. 2. This is shown in Fig. 3. Such three partitions are then fed to its private ANT process. The purpose is to further independently generate a pair of optimal partitions: P_{Fopt} and P_{Sopt} for each sub-domain. P_{Fopt} and P_{Sopt} represent the first and second ranked optimal partitions, respectively. The whole three pairs of P_{Fopt} and P_{Sopt} yielded by three independent ANT processes are all

$P_{rand(1)}$

f_d \ c^k	c^1	c^2
f_1	[0, 0.1)	[0, 0.2)
f_2	[0, 0.6)	[0, 0.2)
f_3	[0, 0.1)	[0, 0.4)

$P_{rand(2)}$

f_d \ c^k	c^1	c^2
f_1	[0.1, 0.7)	[0.2, 0.3)
f_1	[0.6, 0.8)	[0.2, 0.5)
f_1	[0.1, 0.4)	[0.4, 0.9)

$P_{rand(3)}$

f_d \ c^k	c^1	c^2
f_1	[0.7, 1)	[0.3, 1)
f_1	[0.8, 1)	[0.5, 1)
f_1	[0.4, 1)	[0.9, 1)

Fig. 3: The range of possible values for each feature d in cluster c^k with respect to $P_{rand(1)}$, $P_{rand(2)}$ and $P_{rand(3)}$

together compared to each other such that only the two best optimal partitions, P_{FBopt} and P_{SBopt} are chosen. Then, the first-ranked best optimal partition, P_{FBopt} and the second one nearby, P_{SBopt} are set to P_1 and P_2 for continuing redefine the three new sub-domains for all features in the later IFAC iteration.

It is noted that the main loop in algorithm 1 complies with the divide and conquer regulation such that the execution of the ANT processes are performed independently apart from each other based upon the individual sub-domain. Then, all the optimal partitions resulted from the individual ANT seamlessly enter the process of decision making. After termination of the iteration loop, the best one of the resulted optimal partitions is picked up to form a final solution then it is fed to FCM for further refining. The exploration and exploitation are cooperated when the ANT processes are separately functioned upon the three sub-domains. The overall process of IFAC is described:

Algorithm 1: Improved Fuzzy Ant-based Clustering (IFAC)
/* initialization */
1. Normalize all feature values in a range [0, 1]
2. Randomly initialize two partitions of clusters: P_1 and P_2

/* main loop */
3. Repeat
   3.1: Based on P_1 and P_2, the domain of a feature space d, belonging to cluster c^k is divided into three sub-domains as shown in Fig. 2
   3.2: Randomly select three particular partitions such that their feature values are selected in a range of their corresponding sub-domai as demonstrated in Fig. 3
   3.3: Each partition, yielded from 3.2 is fed to its private ANT process, for further independently generating three pairs of optimal partitions one for each sub-domain
   3.4: Compare all together the six optimal partitions, yielded from 3.3 then select only the two best partitions
   3.5: Set those two best optimal partitions to $P_1$and P_2
   Until the criteria of iteration runs is met
4. Feed the best optimal partition, yielded by 3 to FCM for further refining

However, it is noted that the execution cycle of exploitation and exploration is controlled by none of arbitrarily defined parameter. The two mechanisms by themselves are automatically functioned in harmonizing manner. The schemes of such nonparametric controlling techniques signify the important advantage of the IFAC. Although, three sub-domains of a feature space are employed throughout the main loop of the IFAC calculation, the worst-case complexity, big-O of the IFAC relies on the following condition: if N is greater than T then the complexity would be O(DN) else it would be O(DT) where N is the number of sample cases, T is the number of iterations and D refers to number of

dimensions. The other related parameters, e.g., the number of ants as well as number of sub-domains of the search space are counted as small value constants, existing in the clustering process.

EXPERIMENTAL RESULTS

The data sets, tested here consist of two artificial data sets: Artset1 and Artset2 and six well-known data sets, available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/, named Parkinson, Lymphogra-phy, Dermatology, Iris, Contraceptive and Breasttissue.

The real-world data sets with n samples is described as follows: Parkinson (n = 195, d = 22, k = 2) is comprised of the 195 samples are characterized by twenty two features, relating to frequency and noise measures. Two categories exist in the data set, Parkinson's (147 cases) and healthy (48 cases). Lymphography (n = 148, d = 18, k = 4) consists of four different types of lymphatic: normal (2 cases), metastases (81 cases), malign lymph (61 cases) and fibrosis (4 cases). Each type has eighteen features, concerning the structures of lymphatic nodes, for examples. Dermatology (n = 366, d = 34, k = 6) consists of 366 cases, characterized by thirty four features related to erythema and histopathology as well. There are six categories of the samples: psoriasis (112 cases), seboreic dermatitis (61 cases), lichen planus (72 cases), pityriasisrosea (49 cases), cronic dermatitis (52 cases) and pityriasisrubrapilaris (20 cases). Iris (n = 150, d = 4, k = 3) consists of three different species of iris flowers: *Iris setosa*, *Iris versicolour* and *Iris virginica*. For each species, 50 samples with four features: sepal length, sepal width, petal length and petal width were collected. Contraceptive Method Choice (n = 1473, d = 9 and k = 3) is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who either were not pregnant or did not know if they were at the time of interview. The problem is to predict the choice of current contraceptive method. The no use has 629 cases, long-term methods have 334 cases and short-term methods have 510 cases of a woman based on her demographic and socioeconomic characteristics. Breast tissue (n = 106, d = 9 and k = 6) consists of 106 sample cases. The breast tissue information is given via impedance measurement. There are six categories in this data set: carcinoma (21 cases), fibro-adenoma (15 cases), mastopathy (18 cases), glandular (16 cases), connective (14 cases) and adipose (22 cases).

The artificial data set is also provided here. Artset1 (n = 900, d = 2 and k = 3) this is an artificial data set. It is a two-featured problem with three unique classes. A total of 900 patterns are drawn from three independent bivariate normal distributions where classes are distributed according to:

$$N_2\left(\mu=\begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}\right), \sum\begin{bmatrix}0.080 & 0.076\\ 0.076 & 0.074\end{bmatrix},\ i=1, 2, 3$$

$$\mu_{11}=0.163, \mu_{12}=0.147, \mu_{21}=0.535,$$

$$\mu_{22}=0.477, \mu_{31}=0.799$$

Where:
(μ_{i1}/μ_{i2}) = Mean vector of class i
$\sum$ = Covariance matrix

The data set Artset1 is illustrated in Fig. 4a. Artset2 (n = 300, d = 3 and k = 3) this artificial data set is a three-featured problem with three classes and 300 patterns where the sample cases in each class is distributed in such a following manners:

$$\text{Class1} \sim \text{Uniform}\begin{bmatrix}0.512 & 0.798\\ 0.143 & 0.547\\ 0.448 & 0.644\end{bmatrix}$$

$$\text{Class2} \sim \text{Uniform}\begin{bmatrix}0 & 0.275\\ 0 & 0.490\\ 0 & 0.247\end{bmatrix}$$

$$\text{Class3} \sim \text{Uniform}\begin{bmatrix}0.708 & 1\\ 0.461 & 1\\ 0.658 & 1\end{bmatrix}$$

The data set Artset2 is illustrated in Fig. 4b. The results of the proposed method, IFAC and the comparative clustering methods: FAC, ANT, FCM alone and some other types of efficient clustering methods, e.g., SOM and AL are evaluated in this section. IFAC and FAC employ 10 ants, running 30 maximum iterations.

The results are refined by 100 FCM runs. ANT employs the same number of ants, running 40 maximum iterations. The 400 iterations are consumed by FCM, SOM and AL. Such number of runs are assigned for fair comparison tests. The quality of the respective clustering approaches are evaluated and compared. Such a quality is measured by the following criteria.

The objective function values of FCM: This is the sum over all the distance from a sample case to all the centers as defined in Eq. 1. Clearly, the smaller the sum is, the higher the quality of clustering is.

The F-measure: This is related with the precision and the recall from the information retrieval (Dalli, 2003; Handl *et al*., 2003). The precision and the recall are defined as:

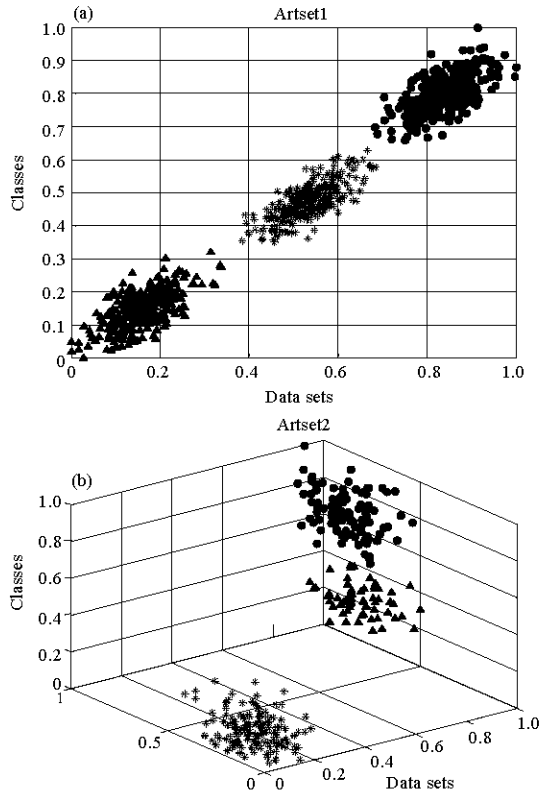


Fig. 4: The artificial data sets

$$p(i,j)=\frac{n_{ij}}{n_i},\ r(i,j)=\frac{n_{ij}}{n_j} \tag{4}$$

Where each class i (given by the class labels of the used data set) is regarded as the set of n_i items desired for a query and each cluster j (generated by the algorithm) is regarded as the set of n_j items retrieved for a query. n_{ij} is the number of sample cases of the class i within cluster j. For a class i and a cluster j, the F-measure is defined as:

$$F(i,j)=\frac{\left(b^2+1\right)\times p(i,j)\times r(i,j)}{b^2\times p(i,j)\times r(i,j)} \tag{5}$$

where researchers choose b = 1 to obtain equal weighting for p(i, j) and r(i, j). The overall F-measure for the data set of size n is given by:

$$F\text{-}\sum_i \frac{n_i}{n_{max_i}}\{F(i,j)\} \tag{6}$$

The bigger the F-measure is, the better the clustering algorithm is.

Xie-Beni index (XB): The XB (Xie and Beni, 1991; Olson, 1995) is called the compactness and separation validity function as shown in Eq. 7. The compactness and separation measure are respectively indicated in numerator and denominator of the equation and are defined in Eq. 8 and 9. Small values of XB are expected for compact and well-separated clusters:

$$XB(C,X)=\frac{\sigma(C,X)}{n\times sep(C)} \tag{7}$$

$$\sigma(C,X)=\sum_{K=1}^{K}\sum_{i\in c^k} D^2\left(c^k,x_i\right) \tag{8}$$

$$sep(C)=\min_{i\neq k}\left\|x_i-c^k\right\|^2 \tag{9}$$

Where:

n = The number of sample cases
x_i = Sample case i
K = Total number of clusters
$D^2(c^k, x_i)$ = A Euclidian distance between c^k and x_i
c^k = The center of cluster k

The algorithms of IFAC and all related clustering methods are implemented using MATLAB 7.10 (R2010a) on a CPU 2.4 GHz Core2™Quad with 4 GB RAM. The experiments are performed on the aforementioned eight data sets. The results, yielded by the six related algorithm are averages of 10 independent cross-validation runs. The derived boxplots in Fig. 5 signifies the competitive F-measure degrees of FAC and IFAC in most cases. However, the proficiently low standard deviation of the F-measure degree of IFAC is pointed. Thus, the superiority of IFAC is regarded. Although, the results in Table 1 indicate more runtimes consumed by the IFAC than the others, Fig. 6 exhibits the natural logarithmic values of FCM objective and XB degree of the six algorithms. The means and standard deviations (in parenthesizes) of the IFAC results are remarkable, compared to the others. This confirms the efficiency of the IFAC in terms of both minimum dissimilarity within a cluster and maximum separateness between different clusters. Such distinguishing results of the IFAC are still shown in the cases of high dimensional data sets, e.g., dermatology, lymphography and Parkinson with number of features D = 34, 18 and 22, respectively. Besides, Table 2 displays the results of the IFAC clustering on the eight data sets, relying on various particular numbers of ants employed. One would see most of effective results are generated by using ten ants. Increasing number of ants does not raise the clustering efficiency. Employing such a few numbers of ants denotes the interesting merit of the IFAC.

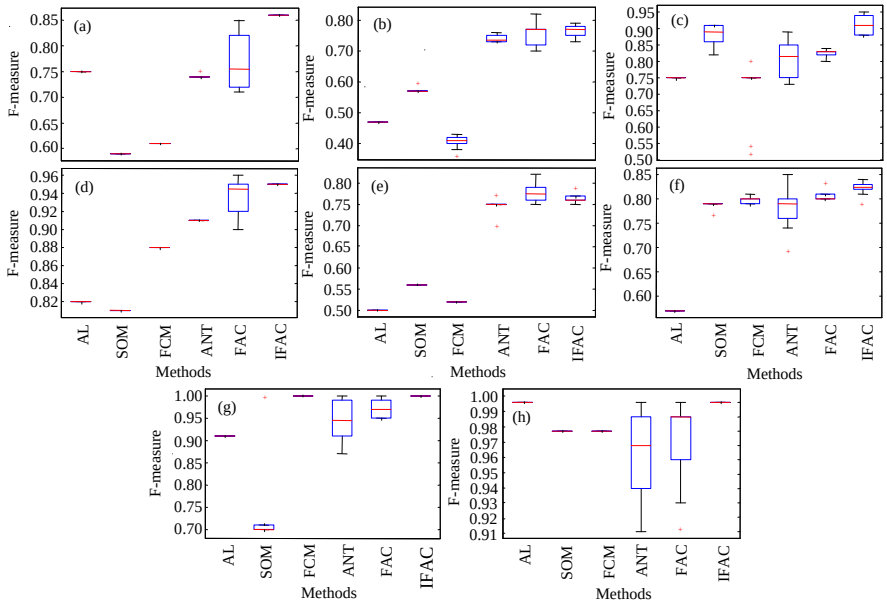


Fig. 5: Ranges of F-measure degrees resulted from running the six clustering algorithms on the eight data sets; a) Parkinson; b) Lymtography; c) Dermatology; d) Iris; e) Contraceptive; f) Breast tissue; g) Artset1; h) Artest2

Table 1: Runtimes in seconds, consumed by the six algorithms on the eight data sets. The means and standard deviations (in parenthesizes) for 10 independent cross-validation runs are reported. Bold face indicates the best runtime

Sources	IFAC	FAC	ANT	FCM	SOM	AL
Parkinson	2.7005 (0.0154)	0.5067 (0.0003)	0.2267 (0.0789)	**0.0102 (0.0147)**	0.9330 (0.0935)	0.1110 (0.2572)
Lymphography	1.1644 (0.1238)	1.0399 (0.0438)	0.1327 (0.0090)	**0.0106 (0.0035)**	0.9520 (0.0782)	0.0225 (0.0028)
Dermatology	1.2970 (0.1622)	1.1509 (0.0616)	0.2653 (0.0148)	**0.0319 (0.0033)**	1.4241 (0.0418)	0.0620 (0.0161)
Iris	3.0340 (0.1052)	0.3351 (0.0328)	0.0462 (0.0022)	**0.0062 (0.0012)**	0.8886 (0.0670)	0.0234 (0.0088)
Contraceptive	2.9716 (0.4551)	2.9044 (0.4690)	2.1069 (0.7012)	**0.2567 (0.0553)**	3.7740 (0.9148)	2.7051 (0.0357)
Breast tissue	1.6281 (0.1800)	1.5991 (0.7973)	1.3630 (0.0762)	**0.0186 (0.0085)**	1.8897 (0.0843)	0.0202 (0.0044)
Artset1	1.0524 (0.0454)	0.3472 (0.0227)	**0.0453 (0.0187)**	0.0590 (0.0386)	0.9232 (0.0328)	0.1333 (0.0135)
Artset2	1.5192 (0.0003)	0.3472 (0.0150)	0.0457 (0.0016)	**0.0148 (0.0041)**	0.9050 (0.0705)	0.1302 (0.0097)

Table 2: Results of IFAC clustering on the eight data sets, relying on various particular numbers of ants employed. The means and standard deviations (in brackets) for 10 independent cross-validation runs are reported. Bold face indicates the best result

Sources	5 ants	10 ants	15 ants	20 ants	40 ants
Parkinson					
F-measure	0.8630 (0.0007)	**0.8630 (0.0003)**	0.8620 (0.0003)	0.8630 (0.0004)	0.8629 (0.0006)
FCM objective values	1.5282 (0.0003)	**1.5192 (0.0003)**	1.5281 (0.0003)	1.5282 (0.0002)	1.5282 (0.0003)
XB	0.0018 (0.0008)	**0.0016 (0.0010)**	0.0018 (0.0007)	0.0020 (0.0007)	0.0018 (0.0009)
Runtime	**2.1336 (0.3680)**	2.7005 (0.0154)	4.9662 (0.5346)	5.5527 (0.0696)	6.7299 (0.4543)
Lymphography					
F-measure	0.7543 (0.0176)	**0.7665 (0.0192)**	0.7610 (0.0216)	0.7591 (0.0279)	0.7648 (0.0245)
FCM objective values	2.5634 (0.0000)	**2.5336 (0.0000)**	2.5634 (0.0001)	2.5634 (0.0001)	2.5634 (0.0001)
XB	**0.0001 (0.0000)**	**0.0001 (0.0000)**	0.0001 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)
Runtime	2.2659 (0.1617)	**1.1644 (0.1238)**	3.1004 (0.1512)	4.0907 (0.0818)	5.5549 (0.2865)
Dermatology					
F-measure	0.9068 (0.0285)	**0.9109 (0.0276)**	0.8991 (0.0205)	0.9073 (0.0302)	0.9103 (0.0302)
FCM objective values	1.0993 (0.0001)	**1.0003 (0.0000)**	**1.0003 (0.0000)**	**1.0003 (0.0000)**	**1.0003 (0.0000)**
XB	0.0005 (0.0002)	0.0004 (0.0005)	0.0002 (0.0001)	**0.0001 (0.0001)**	0.0003 (0.0001)
Runtime	1.5904 (0.3482)	**1.2970 (0.1622)**	3.2524 (0.4763)	4.8235 (0.3654)	5.7185 (1.9720)
Iris					
F-measure	0.9527 (0.0000)	0.9527 (0.0000)	0.9527 (0.0000)	**0.9530 (0.0009)**	0.9527 (0.0000)
FCM objective values	0.5466 (0.0000)	**0.5466 (0.0000)**	0.5466 (0.0001)	0.5466 (0.0001)	0.5466 (0.0000)
XB	0.0514 (0.0001)	**0.0354 (0.0871)**	0.0513 (0.0003)	0.0516 (0.0008)	0.0514 (0.0001)
Runtime	**1.6137 (0.0760)**	3.0340 (0.1052)	2.8929 (0.0472)	3.4637 (0.0259)	4.0115 (0.2039)
Contraceptive					
F-measure	0.7806 (0.0148)	0.7765 (0.0196)	0.7727 (0.0185)	0.7659 (0.0147)	**0.7818 (0.0145)**
FCM objective values	4.6402 (0.0001)	**4.6402 (0.0000)**	**4.6402 (0.0000)**	4.6402 (0.0001)	4.6402 (0.0001)
XB	0.0001 (0.0005)	**0.0001 (0.0000)**	**0.0001 (0.0000)**	0.0001 (0.0001)	0.0002 (0.0000)
Runtime	2.9818 (0.8267)	**2.9716 (0.4551)**	3.8825 (0.2526)	4.7404 (0.9159)	4.5212 (2.4658)

Table 2: Continue

Sources	5 ants	10 ants	15 ants	20 ants	40 ants
Breast tissue					
F-measure	0.8250 (0.0128)	**0.8251 (0.0128)**	0.8250 (0.0128)	0.8250 (0.0128)	0.8251 (0.0128)
FCM objective values	0.0635 (0.0001)	**0.0634 (0.0001)**	0.0635 (0.0001)	0.0635 (0.0001)	0.0635 (0.0001)
XB	**0.0194 (0.0187)**	**0.0194 (0.0187)**	0.0194 (0.0187)	0.0194 (0.0187)	0.0194 (0.0187)
Runtime	**1.3479 (0.2247)**	1.6281 (0.1800)	2.0613 (0.2445)	2.7076 (0.1023)	2.5684 (0.3705)
Artset1					
F-measure	**1.0000 (0.0000)**	**1.0000 (0.0000)**	**1.0000 (0.0000)**	**1.0000 (0.0000)**	**1.0000 (0.0008)**
FCM objective values	0.8378 (0.0001)	0.4666 (0.0458)	**0.4437 (0.0000)**	0.8634 (0.0000)	0.6682 (0.0001)
XB	**0.0540 (0.0063)**	0.0794 (0.0397)	0.0906 (0.0171)	**0.0540 (0.0063)**	0.1023 (0.0542)
Runtime	**1.0306 (0.4914)**	1.0524 (0.0454)	2.3435 (0.0243)	2.2303 (0.1968)	1.6396 (0.1334)
Artset2					
F-measure	0.9999 (0.0008)	**1.0000 (0.0000)**	0.9999 (0.0007)	**1.0000 (0.0008)**	**1.0000 (0.0008)**
FCM objective values	0.5752 (0.0001)	0.6581 (0.0001)	0.6562 (0.0001)	0.6682 (0.0001)	**0.5682 (0.0001)**
XB	**0.1022 (0.0542)**	0.2006 (0.1327)	0.1023 (0.0543)	0.1023 (0.0542)	0.1023 (0.0542)
Runtime	**0.7556 (0.0805)**	1.5192 (0.0003)	1.5457 (0.0425)	1.6396 (0.1334)	2.0038 (0.1342)

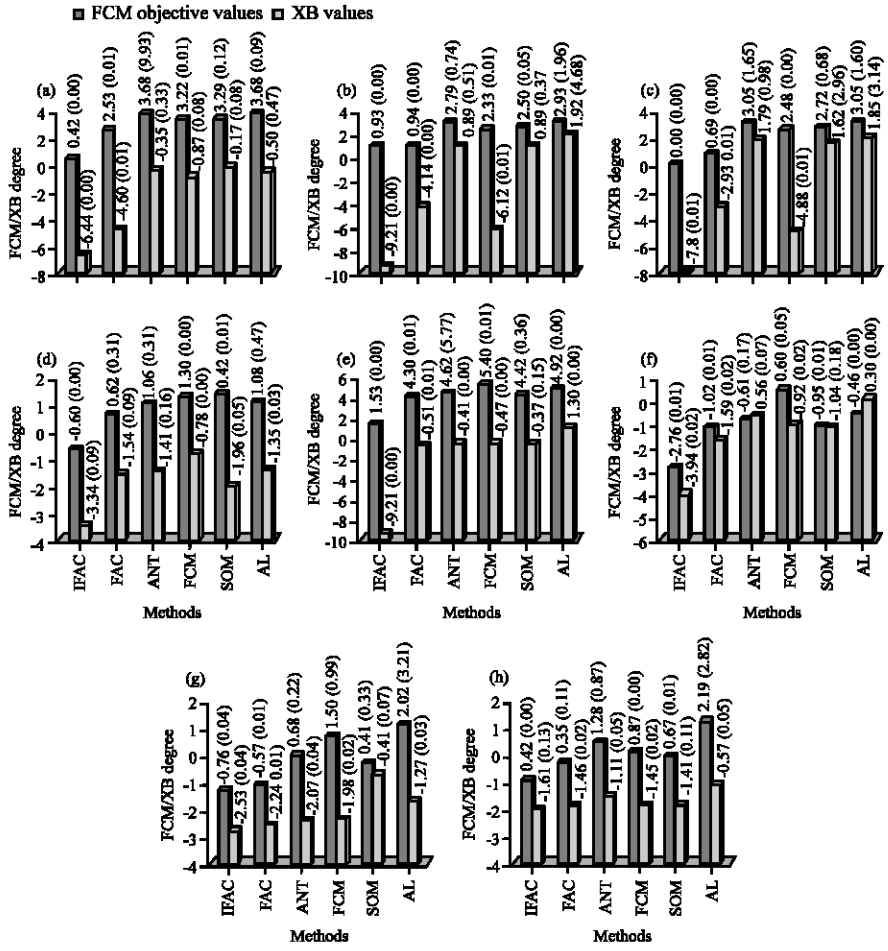


Fig. 6: FCM objective degree and XB values resulted from running the six algorithms on the eight data sets. The mean and standard deviations (in parenthesizes) for 10 independent cross-validation runs are reported on the top of the bars

CONCLUSION

This study presents IFAC which refers to an improve version of fuzzy ant-based clustering. The objective is to apply a nonparametric mechanism to strike the balance between exploration and exploitation during ant-based clustering, aiming to accomplish the global optimal solution. Here, the exploration and exploitation complies

the regulation of divide-and-conquer principle. None of arbitrarily setting parameters is used to control the mechanisms of exploitation and exploration. Such nonparametric mechanisms point the important advantage of the IFAC. The criteria of performance evaluation rely on F-measures, FCM objective degree and Xie-Beni validity index (XB). Additionally, runtimes of the algorithms are provided. The experiments are taken on six benchmarks real-world and two artificial data sets. The comparison tests are performed on the proposed method, IFAC against the former fuzzy ant-based clustering with cluster centroids positioning, ant-based clustering alone as well as some other types of effective clustering algorithms such as SOM and AL. Among all comparative Clustering Methods, the proposed IFAC reports the highest efficiently encouraging results in terms of F-measure, Xie-Beni (XB) validity index and FCM objective degrees. The distinguishing results are also indicated in the cases of high dimensional data sets. The experiments also reveal another merit of the IFAC with respect to the achievement of the powerful clustering using a few numbers of ants. However, the IFAC cannot be applicable when the runtime is quite critical. In the future, modifying the nonparametric methodologies along with finding some other ways of exploration and exploitation techniques will be focused to achieve better runtime results.

REFERENCES

Alpaydm, E., 2004. Introduction to Machine Learning. The MIT Press, Cambridge, MA., pp: 1-3.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. 1st Edn., Plenum Press, New York, USA.

Bonabeau, E., Dorigo, M. and G. Theraulaz, 1999. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York, ISBN-13: 9780195131598, Pages: 307.

Bonham, C.R., 1999. An investigation of exploration and exploitation within cluster oriented genetic algorithms (COGAs). Proceedings of the Genetic and Evolutionary Computation Conference, August 17, 1999, Morgan Kaufmann, pp: 1491-1497.

Chang, H.S., 2004. An ant system based exploration-exploitation for reinforcement learning. Proceedings of the IEEE Conference on Systems, Man and Cybernetics, Volume 1, October 10-13, 2004, IEEE Press, Piscataway, NJ, USA., pp: 3805-3810.

Dalli, A., 2003. Adaptation of the F-measure to cluster-based Lexicon quality evaluation. Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable? February 13, 2003, Budapest, pp: 51-56.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cyber., 3: 32-57.

Fang, C., J. Lee and M.A. Schilling, 2009. Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. Org. Sci., 21: 625-642.

Gan, G., J. Wu and Z. Yang, 2009. A genetic fuzzy k-modes algorithm for clustering categorical data. Expert Syst. Appli., 36: 1615-1620.

Gong, W., Z. Cai, C.X. Ling and J. Du, 2009. Hybrid differential evolution based on fuzzy C-means clustering. Proceedings of the Genetic and Evolutionary Computation Conference, July 8-12, 2009, ACM Press, Montreal, Quebec, Canada, pp: 523-530.

Handl, J., J. Knowles and M. Dorigo, 2003. On the performance of ant-based clustering. Proceedings of the Design and Application of Hybrid Intelligent Systems, November 2003, IOS Press, Australia, pp: 204.

Hastie, T., R. Tibshirani and J. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edn., Springer, New York, pp: 520-528.

Herrero, A., E. Corchado and J. Alfredo, 2011. Unsupervised neural models for country and political risk analysis. Expert Syst. Appli., 38: 13641-13661.

Izakian, H. and A. Abraham, 2011. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Syst. Appl., 38: 1835-1838.

Kanade, P.M. and L.O. Hall, 2003. Fuzzy ants as a clustering concept. Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society, July 24-26, 2003, New York, pp: 227-232.

Kanade, P.M. and L.O. Hall, 2007. Fuzzy ants and clustering. IEEE Trans. Syst. Man Cybernet. A Syst. Humans, 37: 758-769.

Kao, Y., J.C. Lin and S.C. Huang, 2008. Fuzzy clustering by differential evolution. Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, November 26-28, 2008, Kaohsiung, pp: 246-250.

Kiang, M.Y., 2003. A comparative assessment of classification methods. Decision Support Syst., 35: 441-454.

Kohonen, T., 1995. Self-Organizing Maps. Vol. 30, Springer, Berlin, Germany.

Li, D. and C. Yeh, 2008. A non-parametric learning algorithm for small manufacturing data sets. Expert Syst. Appli., 34: 391-398.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, January 17-20, 1967, Berkeley, CA., USA., pp: 281-297.

Olson, C.F., 1995. Parallel algorithms for hierarchical clustering. Parallel Comput., 21: 1313-1325.

Pai, D.R., K.D. Lawrence, R.K. Klimberg and S.M. Lawrence, 2012. Experimental comparison of parametric, non-parametric and hybrid multigroup classification. Expert Syst. Appli., 39: 8593-8603.

Rojas, R., 1996. Neural Networks. Springer-Verlag, Berlin, Germany.

Rugina, R. and M. Rinard, 2001. Recursion unrolling for divide and conquer programs. Proceedings of the 13th International Workshop on Languages and Compilers for Parallel Computing, August 10-12, 2000, Yorktown Heights, NY., USA., pp: 34-48.

Supratid, S. and P. Julrode, 2011. A performance comparison using principal component analysis and differential evolution on fuzzy c-means and k-harmonic means. Rangsit J. Arts Sci., 1: 127-137.

Tan, P., M. Steinbach and V. Kumar, 2004. Introduction to Data Mining. Pearson Addison Wesley, Upper Saddle River, NJ.

Tian, Y., D. Liu and H. Qi, 2009. K-harmonic means data clustering with differential evolution. Proceedings of the International Conference on Future BioMedical Information Engineering, December 13-14, 2009, Sanya, pp: 369-372.

Webb, A.R., 2002. Statistical Pattern Recognition. John Wiley and Sons Ltd., USA.

Xie, X.L. and G. Beni, 1991. A validity Measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell., 13: 841-847.