Using nonparametric balance between exploitation and exploration on fuzzy ant-based clustering

Phichete Julrode and Siriporn Supratid

***Abstract*—**Balancing between exploitation and exploration in a search space is significant key to cope with fast convergence and divergence. Thus, such type of balance can lead to achieve the global optimal solution. This paper proposed a new improved version of fuzzy ant-based clustering, using a nonparametric balance between exploitation and exploration (NIFAC). Such proposed method does not only integrate ant-based clustering and fuzzy c-means, but also remarkably apply the techniques of balancing between exploitation and exploration without using any arbitrarily defined parameter to control the search in the balance. The performance measurements relate to F-measures, FCM objective degree and Xie-Beni validity index. The experiments are operated on real-world as well as artificial data sets. The results show the prominent higher performance of the proposed method in terms of clustering correctness than several other types of effective methods including other related ant-based clustering and fuzzy means.

***Keywords*—**Fuzzyc-means,Ant-basedclustering, Divide-and-conquer, Nonparametric, Exploration, Exploitation.

I. Introduction

CLUSTERINGis one of the most important unsupervised learning techniques [1], [2]. 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 [3], data mining [4], machine learning [5], etc. For solving clustering problems, efficient approaches such as self-organizing feature maps (SOM) [6], Average linkages (AL) [7] have been successfully applied. On the other side, K-means [8], a partitional type of clustering employs simply basic idea relating to find cluster centers, then refining them [9]. Unlike SOM, k-means and AL, fuzzy c- means (FCM) [10], [11] which is the soft clustering version of k-means 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 are proposed, aiming to integration between FCM and powerful evolutionary optimization algorithms, i.e. a combination algorithm between particle swarm optimization (PSO) [12] and FCM [13], [14] as well as differential evolution (DE), K-Harmonic means (KHM) and FCM [15], [16], [17], [18]. Ant-based algorithm has been developed using swarm intelligence principles that emphasize distributiveness, direct or indirect interactions among relatively simple agents, flexibility, and robustness [19]. By such competent characteristics, ant-based clustering more relieves the fast convergence during searching process than several other evolutionary approaches [20]. Fuzzy ant-based clustering was primarily proposed by Kanade and Hall [21]. 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 different ones; then the cluster centers, found by the ants are refined, using FCM. In later version [22], 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, FCM is subsequently applied next to the ant clustering in order to achieve better cluster results. However, ant clustering which exploits the search space around the optimally best solution may not get globally best solution. Vice versa, the clustering which explores the search space possibly gets better solution by enhancing the diversity of solutions, but needs more time to converge. Therefore it is necessary to strike the balance between exploration and exploitation for achieving globally best solution [23], [24], [25]. Several methods attempt to accomplish such an equilibration. Most of them, related to parametric methods that may lead to a biased and overly optimistic clustering process; thus limit the usefulness of the model [26]. On the other side, a non-parametric algorithm automatically adjusts the learning by the algorithm itself; none of arbitrarily setting parameters is used to control or direct the algorithm functions. The nonparametric learning algorithm, proposed by [27] is employed to speed up stabilizing the learning task; and can dynamically improve deriving

P. Julrode is with the Informatics Department of Science and Technology, Phuket Rajabhat University, Phuket 83000 Thailand (phone: +667-621-1959; fax: +667-621-1778; e-mail: phichete@ pkru.ac.th).

S. Supratid is with the Information Technology Department, University of Rangsit, Pathum Thani, 12000 Thailand (e-mail: siri_sup1@hotmail.com).

knowledge. Another non-parametric algorithm combines learning technique and a linear programming approach [28], [29], [30]; this combination considerably improves the classification accuracy as well as reliability.

This paper, thereby proposes a new improved version of fuzzy ant-based clustering, using a nonparametric balance between exploitation and exploration (NIFAC). Here, the exploration and exploitation complies the regulation of divide-and-conquer principle [31]. The improvement emphasizes on the nonparametric balance between exploitation and exploration search techniques during the ant process. Here, the nonparametric balance of exploitation and exploration is remarkable, according to this work such that none of arbitrarily setting parameters is used to control the mechanisms of exploitation and exploration. The criteria of performance evaluation, applied here rely on F-measures, FCM objective degree and Xie-Beni validity index (XB); additionally, runtime 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, NIFAC against fuzzy ant-based clustering, ant-based clustering, as well as some other types of effective clustering algorithms such as SOM and AL.

The rest of the paper is organized as follows. In section 2, fuzzy ant-based clustering is reviewed. Section 3 presents the proposed algorithm, NIFAC. Section 4 reports the experimental results. Finally, section 5 makes conclusion of this work.

II. Fuzzy Ant-Based Clustering with Cluster Centroids Positioning (FAC)

The fuzzy ant-based clustering with cluster centroids positioning, FAC formally proposed by [21] is a combination between ant-based clustering (ANT) and FCM aims to search for optimal partition of centers of clusters. Initially, the feature values are normalized between 0 and 1. For all feature spaces in every cluster partitions, an ant is assigned to a particular feature of a cluster in a partition. The ants never change the feature, cluster or partition assigned to them. To search for the new partition, ants randomly move the clusters with corporative style.

There are two directions 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):

$$FCM\_ObjectiveFunction = \sum_{i=1}^{N}\sum_{k=1}^{K} \mu_{ik} \left\| \mathbf{x}_i - \mathbf{c}^k \right\|^2 \qquad (1)$$

where μ_{ik} represents membership of x_i, sample i in cluster $\mathbf{c}^k$; for crisp data, μ_{ik} is zero if x_i is in cluster $\mathbf{c}^k$; and is one if x_i is not.

III. A New Improved Version of Fuzzy Ant-Based Clustering (NIFAC)

Relating to ANT and even FAC, it is still highly possible that ants may exploit the search space around the optimal best solution only. Thus, globally best solution cannot be achieved. Therefore, a new improved version of fuzzy ant-based clustering, using a nonparametric balance between exploitation and exploration (NIFAC) is proposed in this paper. The objective is to accomplish the nonparametric balance between exploitation and exploration operation aiming to obtain globally optimal partition of centers of clusters. The functions of exploration and exploitation follow the principle of divide and conquer, that would be shortly explained. Additionally, none of arbitrarily parameters is defined to control the operations between both types of searches. The overall process of NIFAC is delineated in Fig.1.

In the beginning of the algorithm of NIFAC described in Fig. 1, the data are normalized in a range [0, 1], at the initial step. The two initial partitions, P_1 and P_2 are randomly selected and respectively represented by the matrices: $[f_d(\mathbf{c}^k_{P_1})]_{D\times K}$ and $[f_d(\mathbf{c}^k_{P2})]_{D\times K}$. Such matrices are composed of feature d in cluster $\mathbf{c}^k$, $d = 1, \ldots, D$ and $k = 1,\ldots, K$ 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 NIFAC. The domain of an individual feature space d, belonging to cluster $\mathbf{c}^k$ in each partition P_1 and P_2 is divided into three sub-domains, defined by $f_{d(1)}(\mathbf{c}^k)$, $f_{d(2)}(\mathbf{c}^k)$ and $f_{d(3)}(\mathbf{c}^k)$. Such three sub-domains are illustrated in Fig.2 $f_d^{low}(\mathbf{c}^k)$ and $f_d^{high}(\mathbf{c}^k)$ refer to boundaries, 'low' and 'high' associated with those sub-domains. Fig. 3 delineates an example of finding such boundaries and three sub-domains. Partition P_1 and $P_2$are comprised of two clusters, $\mathbf{c}^1$ and $\mathbf{c}^2$; 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 $f_d^{low}(\mathbf{c}^k)$ and the larger one is designated $f_d^{high}(\mathbf{c}^k)$. This is supported by Eq. (2) and (3).

$$f_d^{low}(\mathbf{c}^k) = \min\left(f_d(\mathbf{c}^k_{P_1}), f_d(\mathbf{c}^k_{P_2})\right) \qquad (2)$$

$$f_d^{high}(\mathbf{c}^k) = \max\left(f_d(\mathbf{c}^k_{P_1}), f_d(\mathbf{c}^k_{P_2})\right) \qquad (3).$$

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.3. Two gray areas of sub-domains respectively refer to $f_{d(1)}(\mathbf{c}^k)$ and $f_{d(3)}(\mathbf{c}^k)$, relevant to Fig. 2. The white area of sub-domain relates to $f_{d(2)}(\mathbf{c}^k)$. Such $f_{d(2)}(\mathbf{c}^k)$ covers the space in between the corresponding clusters in the two

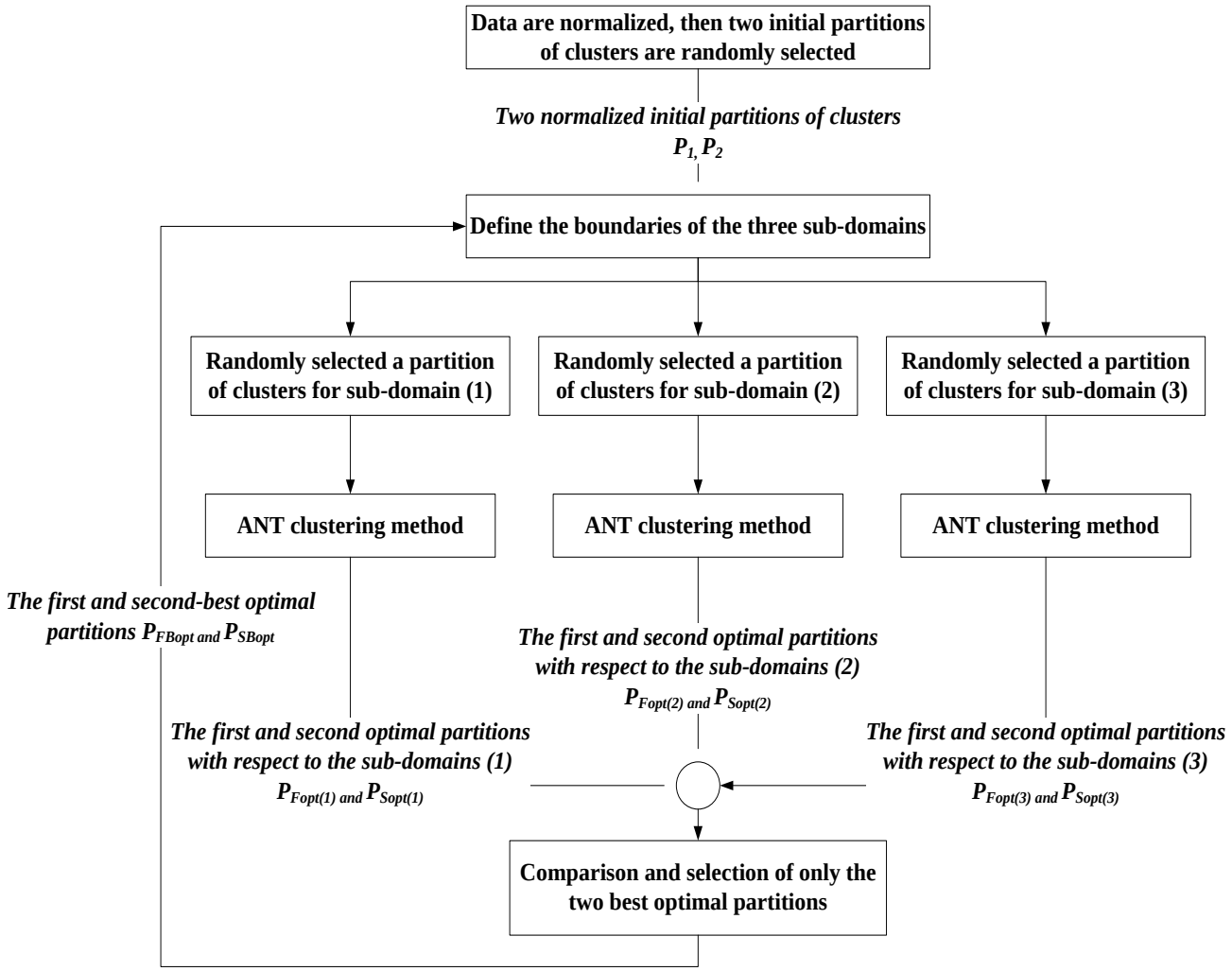


Fig. 1 The overall process of the new improved version of fuzzy ant-based clustering (NIFAC)

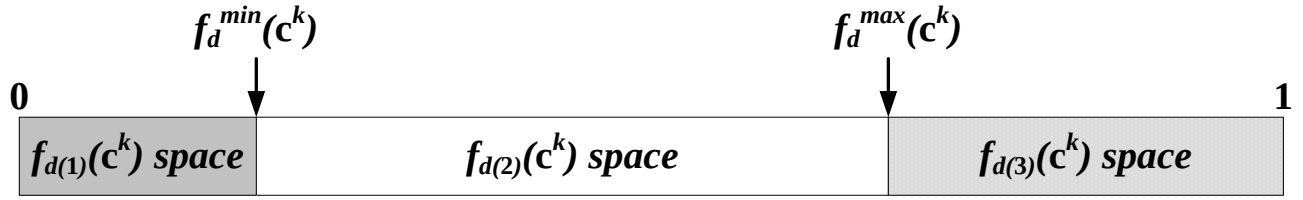


Fig. 2 An individual $f_d(\mathbf{c}^k)$ feature space is divided into three sub-domains

best optimal partitions.Hence, such sub-domain $f_{d(2)}(\mathbf{c}^k)$ can be regarded as an exploitation space. Whereas,theothertwosub-domains $f_{d(1)}(\mathbf{c}^k)$ and $f_{d(3)}(\mathbf{c}^k)$ 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 NIFAC. 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. 3. This is shown in Fig.4. 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 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 NIFAC iteration.

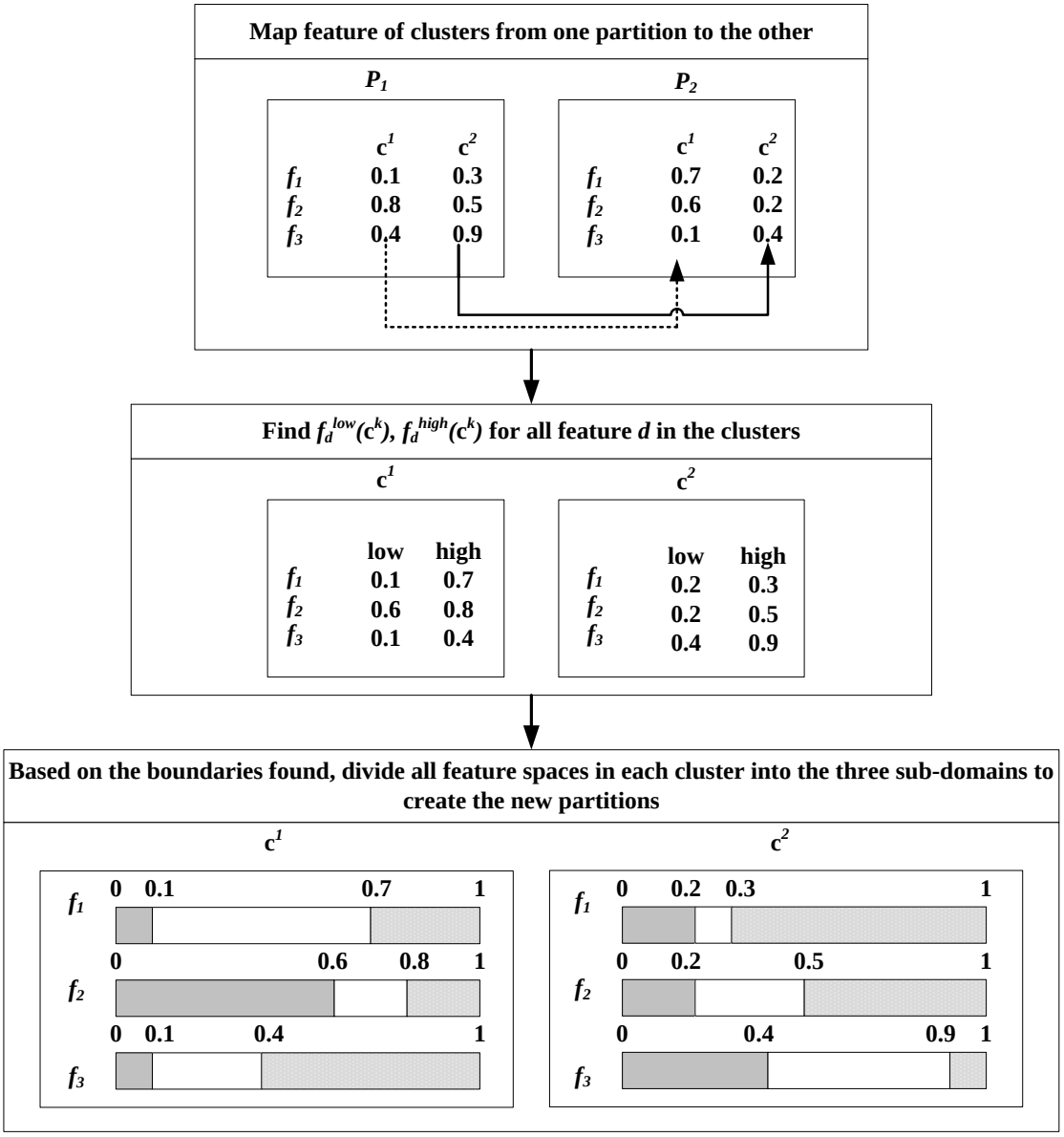


Fig.3 An example of finding the boundaries $f_d^{low}(\mathbf{c}^k)$ and $f_d^{high}(\mathbf{c}^k)$ of the threesub-domains

$P_{rand(1)}$

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_2	[0.6, 0.8)	[0.2, 0.5)
f_3	[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_2	[0.8, 1]	[0.5, 1]
f_3	[0.4, 1]	[0.9, 1]

Fig.4 The range of possible values for each feature *d* in cluster $\mathbf{c}^k$ with respect to $P_{rand(1)}$, $P_{rand(2)}$ and $P_{rand(3)}$

Algorithm 1: A new improved version of fuzzy ant-based clustering (NIFAC)

/* initialization */

1. Normalize all feature values in a range [0, 1]
2. Randomly initialize two partitions of clusters

/* main loop */

3. Repeat
   - 3.1 Define the boundaries $\left[0, f_d^{min}(\mathbf{c}^k)\right), \left[f_d^{min}(\mathbf{c}^k), f_d^{max}(\mathbf{c}^k)\right]$ and $\left(f_d^{max}(\mathbf{c}^k), 1\right]$ of three sub-domain, $f_{d(i)}(\mathbf{c}^k)$, $d = 1, \ldots, D$ and $k = 1, \ldots, K$, i=1,…,3 where i represent an index of number of sub-domain using Eq. (2) and (3)
   - 3.2 Initialize a partitions of clusters with respect each of the three sub-domain $f_{d(i)}(\mathbf{c}^k)$ from 3.1
   - 3.3 Call ant-based clustering (ANT) method to generate new six clusters: $f_d(\mathbf{c}^k_{P_{Fopt}})$ and $f_d(\mathbf{c}^k_{P_{Sopt}})$, the indices P_{Fopt} and P_{Sopt} respectively represents the first and the second optimal partitions, belonging to a particular of sub-domain
   - 3.4 From optimal partitions yielded in 3.3 compare and select of first and second best optimal partitions consecutively contains feature $f_d\left(\mathbf{c}^k_{P_{FBopt}}\right)$ and $f_d\left(\mathbf{c}^k_{P_{SBopt}}\right)$:

   Until the criteria of iteration runs is met

/* FCM */

4. The first-best partition of clusters would be brought into FCM
5. Perform FCM clustering to obtain the final best partition

Fig. 5 The algorithm of NIFAC

It is noted that the main loop in Algorithm 1 in Fig. 5 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. In Fig. 5, the overall process of NIFAC is described. One would see that the exploration and exploitation mechanisms, according to NIFAC iteration runs follow the divide-and-conquer principle such that a feature space $f_d(\mathbf{c}^k)$ is divided into several different sub-domains; afterwards, a pair of optimal partition solutions is yielded independently from each of those sub-domains using ANT; then, all of them seamlessly enter the process of decision making; eventually, the best one of all optimal partitions is picked up to form a completely final solution.

Moreover, it is obvious that the cycle of exploitation and exploration executions is controlled by none of arbitrarily defined parameter. The schemes of such nonparametric mechanisms specify the important advantage of NIFAC. Although three regions of a feature space are involved within the search calculation, the worst-case complexity, big-O of NIFAC 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 related other 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.

IV. Experimental Results

A. Data Sets

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, Lymphography, Dermatology, Iris, Contraceptive and Breasttissue. The detail of those data sets is described as follows:

Parkinson (n = 195, d = 22, k = 2), which consists of 195 objects characterized by twenty-two features: average vocal fundamental frequency, maximum vocal fundamental

frequency, minimum vocal fundamental frequency, five several measures of variation in fundamental frequency, six several measures of variation in amplitude, two measures of ratio of noise to tonal components in the voice, two nonlinear dynamical complexity measures , three nonlinear measures of fundamental frequency variation and signal fractal scaling exponent. There are two categories in the sample cases: Parkinson's (147 cases) and healthy (48 cases).

Lymphography (n = 148, d = 18, k = 4), which consists of four different types of lymphatic: normal find (2 cases), metastases (81 cases), malign lymph (61 cases), and fibrosis (4 cases). Each type has eighteen features, which are lymphatic's, Block_of_ affere, Bl_of_lymph_c, Bl_of_lymph_s, By_pass, Extravasates, Regeneration_of, Early_uptake_in, Lyp_nodes_ dimin, Lym_nodes_enlar, Changes_in_lym, Defect_in_node, Changes_in_node, Change_in_stru, Special_forms, Dislocation_of, Exclusion_of_no, No_of_nodes_in.

Dermatology (n = 366, d = 34, k = 6), which consists of 366 cases characterized by thirty-four features: erythema, scaling, definite borders, itching, koebner phenomenon, polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, family history, age, melanin incontinence, eosinophils in the infiltrate, PNL infiltrate, fibrosis of the papillary dermis, exocytosis, acanthosis, hyperkeratosis, para keratosis, clubbing of the rete ridges, elongation of the rete ridges, thinning of the suprapapillary epidermis, spongiform pustule, munromicro-abcess, focal hypergranulosis, disappearance of the granular layer, vacuolisation and damage of basal layer, spongiosis, saw-tooth appearance of retes, follicular horn plug, perifollicularparakeratosis, inflammatory monoluclear infiltrate and band-like infiltrate. There are six categories in the: psoriasis (112 cases), seboreic dermatitis (61 cases), lichen planus (72 cases), pityriasisrosea (49 cases), cronic dermatitis (52 cases) and pityriasis-rubrapilaris (20 cases).

Iris (n = 150, d = 4, k = 3), which 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, k = 3): This dataset 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 (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 socioeco-nomic characteristics.

Breast Tissue (n = 106, d = 9, k = 6): This data set, consisting of 106 sample casesis characterized by 9 such features as impedivity (ohm) at zero frequency (IO), phase angle at 500 KHz, high-frequency slope of phase angle, impedance distance between spectral ends (DA), area under spectrum, area normalized by DA, maximum of the spectrum, distance between IO and real part of the maximum frequency point and length of the spectral curve. There are six categories in the data set: carcinoma (21 cases), fibro-adenoma (15 cases), mastopathy (18 cases), glandular (16 cases), connective (14 cases) and adipose (22 cases).

Artset1 (n = 900, d = 2, 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 $N_2\left(\mu = \begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}, \Sigma\begin{bmatrix}0.080 & 0.076\\ 0.076 & 0.074\end{bmatrix}\right), i = 1,2,3,\ \mu_{11} = 0.163,\ \mu_{12} = 0.147,\ \mu_{21} = 0.535,\ \mu_{22} = 0.477, \mu_{31} = 0.838,\ \mu_{32} = 0.799,$ where $\begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}$ are mean vector of class i and Σ are covariance matrix, respectively. The data set Artset1 is illustrated in Fig. 6

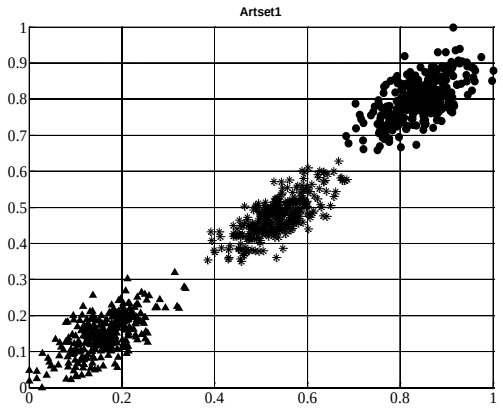


Fig. 6. The Artset1 of artificial data sets.

Artset2 (n = 300, d = 3, 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 manner; Class1 ~ Uniform $\begin{bmatrix}0.512 & 0.798\\ 0.143 & 0.547\\ 0.448 & 0.644\end{bmatrix}$, Class2 ~ Uniform $\begin{bmatrix}0 & 0.275\\ 0 & 0.490\\ 0 & 0.247\end{bmatrix}$, Class3 ~ Uniform $\begin{bmatrix}0.708 & 1\\ 0.461 & 1\\ 0.658 & 1\end{bmatrix}$. The data set Artset2 is illustrated in Fig.7

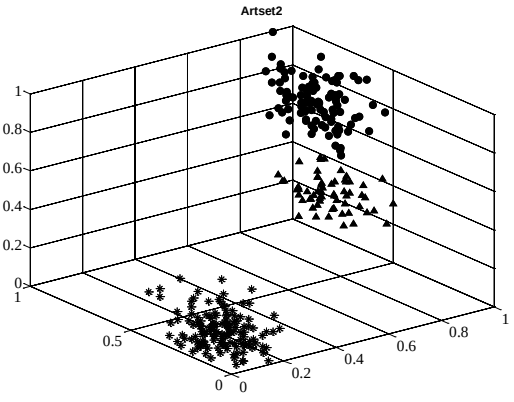


Fig. 7. The Artset2of artificial data sets.

B. Experimental results

The results of the proposed method, NIFAC and other related, FAC, FCM alone and some other types of efficient clustering methods, e.g. SOM and AL are evaluated in this section. Both NIFAC and FAC employ 10 ants, 30 maximum iteration runs for each ant. Their results are refined by 100 FCM runs. 400 iterations are consumed by FCM, SOM and AL. Such criteria are defined for fair comparison tests. The quality of the respective clustering are evaluated and compared. Such a quality is measured by the following criteria

C. 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.

D. The F-Measure

This is related with the precision and the recall from the information retrieval [32], [33]. The precision and the recall are defined as:

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

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:

$$F(i, j) = \left(b^2+1\right)\cdot p(i, j)\cdot r(i, j)/b^2\cdot p(i, j)\cdot r(i, j) \quad (5)$$

where we 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

$$F = \sum_i n_i/N \; max_j \{F(i, j)\} \quad (6)$$

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

E. Xie-Beni index (XB)

The XB [34], [35] 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.

$$XB(\mathbf{C}, \mathbf{X}) = \sigma(\mathbf{C}, \mathbf{X})/N \times sep(\mathbf{C}) \quad (7)$$

$$\sigma(\mathbf{C}, \mathbf{X}) = \sum_{k=1}^{K} \sum_{i\in \mathbf{c}^k} D^2(\mathbf{c}^k, \mathbf{x}_i) \quad (8)$$

$$sep(\mathbf{C}) = \min_{i \neq k} \left\| \mathbf{x}_i - \mathbf{c}^k \right\|^2 \quad (9)$$

where N is the number of sample cases; $\mathbf{x}_i$ is sample case i; K is total number of clusters; $D^2(\mathbf{c}^k, \mathbf{x}_i)$ represents a Euclidian distance between $\mathbf{c}^k$ and $\mathbf{x}_i$, where $\mathbf{c}^k$ represents cluster center.

The algorithms of NIFAC 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 experimental results are averages of 10 runs of simulation. Table I summarizes F-measure, XB and FCM objective degree results of NIFAC, FAC, FCM, SOM and AL; moreover, runtime of the algorithms are provided. The figure in the brackets shows the standard deviations for 10 independent runs. For nearly all the data sets, the proposed method NIFAC yields higher F-measure results than other comparative clustering methods. Most of the best runtime is generated by FCM. Although NIFAC consumes more runtimes than the others, the remarkable superiority of NIFAC is revealed with regard of XB and FCM objective degree. This confirms the efficiency of NIFAC in terms of both minimum dissimilarity within a cluster and maximum separateness between different clusters. Such distinguishing results of NIFAC are also 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 II displays the quality of NIFAC, with regard to various numbers of ants. One can see most of effective results are generated by using ten ants. Increasing number of ants does not raise the clustering efficiency, in most cases. This denotes the interesting merit of NIFAC.

V. Discussion

It is seen from the experimental results in Table I, Fig. 8 and 9 the superiority of the proposed NIFAC over the comparative methods: FAC, ANT, FCM, SOM and AL for all data sets. In Table I, most of data sets show close F-measure values of NIFAC and FAC. However, NIFAC has notable higher value of F-measure than FAC and ANT. F-measure generated by FCM, SOM and AL for most data sets are not significant. Fig. 8 and 9 exhibit most of the remarkable surpassingness of NIFAC over the other methods in terms of natural logarithms of FCM objective and XB degree consecutively. FCM objective degrees, averaged for all data sets relating NIFAC, FAC, ANT, FCM, SOM and AL are 1.4285, 11.9026, 23.4680, 35.1659, 17.7721 and 29.6555 respectively; while XB values are 0.0421, 0.1798, 1.3841, 0.2855, 1.3266 and 2.4905. FCM and XB degree produced by NIFAC are 87.998% and 76.585% better than the second best clustering methods, FAC. Nevertheless, the best runtime consumption belongs to FCM for seven out of eight data sets.Considering Table II, using 10 ants based on NIFAC yields the best F-measure values on six high dimensional data sets out of all the eight data sets; whereas using 40 ants accomplishes on four of the eight. The derived boxplots in Fig. 10 signifies the competitive F-measure degrees of FAC and NIFAC in most cases. However, the proficiently low standard deviation of the F-measure degree of NIFAC is pointed. Thus, the superiority of NIFAC is regarded.

Table I Results of NIFAC, FAC, ANT, FCM, SOM and AL clustering on eight data sets. The quality of clustering is evaluated using F-Measure. Runtimes (seconds) are additionally provided. The table shows the means and standard deviations (in brackets) for 10 independent cross-validation runs. Bold face indicates the best result out of the six algorithms.

Source	*NIFAC*	*FAC*	*ANT*	*FCM*	*SOM*	*AL*
Parkinson						
F-Measure	**0.8630(0.0003)**	0.7665(0.0539)	0.7440(0.0012)	0.6143(0.0000)	0.5921(0.0020)	0.7522(0.0000)
Runtime	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						
F-Measure	**0.7665(0.0192)**	0.7556(0.0426)	0.7396(0.0089)	0.6053(0.0165)	0.5739(0.0022)	0.5733(0.0000)
Runtime	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						
F-Measure	**0.9109(0.0276)**	0.8294(0.0089)	0.8023(0.0583)	0.7086(0.0985)	0.8835(0.0306)	0.7483(0.0000)
Runtime	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						
F-Measure	**0.9527(0.0000)**	0.9370(0.0192)	0.9065(0.0000)	0.8797(0.0000)	0.8111(0.0000)	0.8153(0.0000)
Runtime	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						
F-Measure	**0.7765(0.0196)**	0.7529(0.0196)	0.7444(0.0196)	0.6208(0.0003)	0.5582(0.0015)	0.5051(0.0000)
Runtime	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						
F-Measure	**0.8251(0.0127)**	0.8064(0.0084)	0.7795(0.0412)	0.7979(0.0041)	0.7885(0.0040)	0.5654(0.0001)
Runtime	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						
F-Measure	**1.0000(0.0000)**	**1.0000(0.0000)**	0.9723(0.0219)	**1.0000(0.0000)**	0.7630(0.1244)	0.9145(0.0000)
Runtime	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						
F-Measure	**1.0000(0.0000)**	0.9998(0.0000)	0.9731(0.0297)	0.9804(0.0000)	0.9804(0.0000)	0.9970(0.0000)
Runtime	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 II Results of NIFAC clustering on eight data sets, depending on various particular numbers of ants used. The quality of clustering is evaluated using F-Measure. Runtimes (seconds) are additionally provided. The table shows the means and standard deviations (in brackets) for 10 independent cross-validation runs. Bold face indicates the best result, related to a particular number of ants used.

Source	*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)
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)
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)
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)
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)**
Runtime	2.9818 (0.8267)	**2.9716 (0.4551)**	3.8825 (0.2526)	4.7404 (0.9159)	4.5212 (2.4658)

Table II (Cont.) Results of NIFAC clustering on eight data sets, depending on various particular numbers of ants used. The quality of clustering is evaluated using F-Measure. Runtimes (seconds) are additionally provided. The table shows the means and standard deviations (in brackets) for 10 independent cross-validation runs. Bold face indicates the best result, related to a particular number of ants used.

Source	*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)**
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)**
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)**
Runtime	**0.7556 (0.0805)**	1.5192 (0.0003)	1.5457 (0.0425)	1.6396 (0.1334)	2.0038 (0.1342)

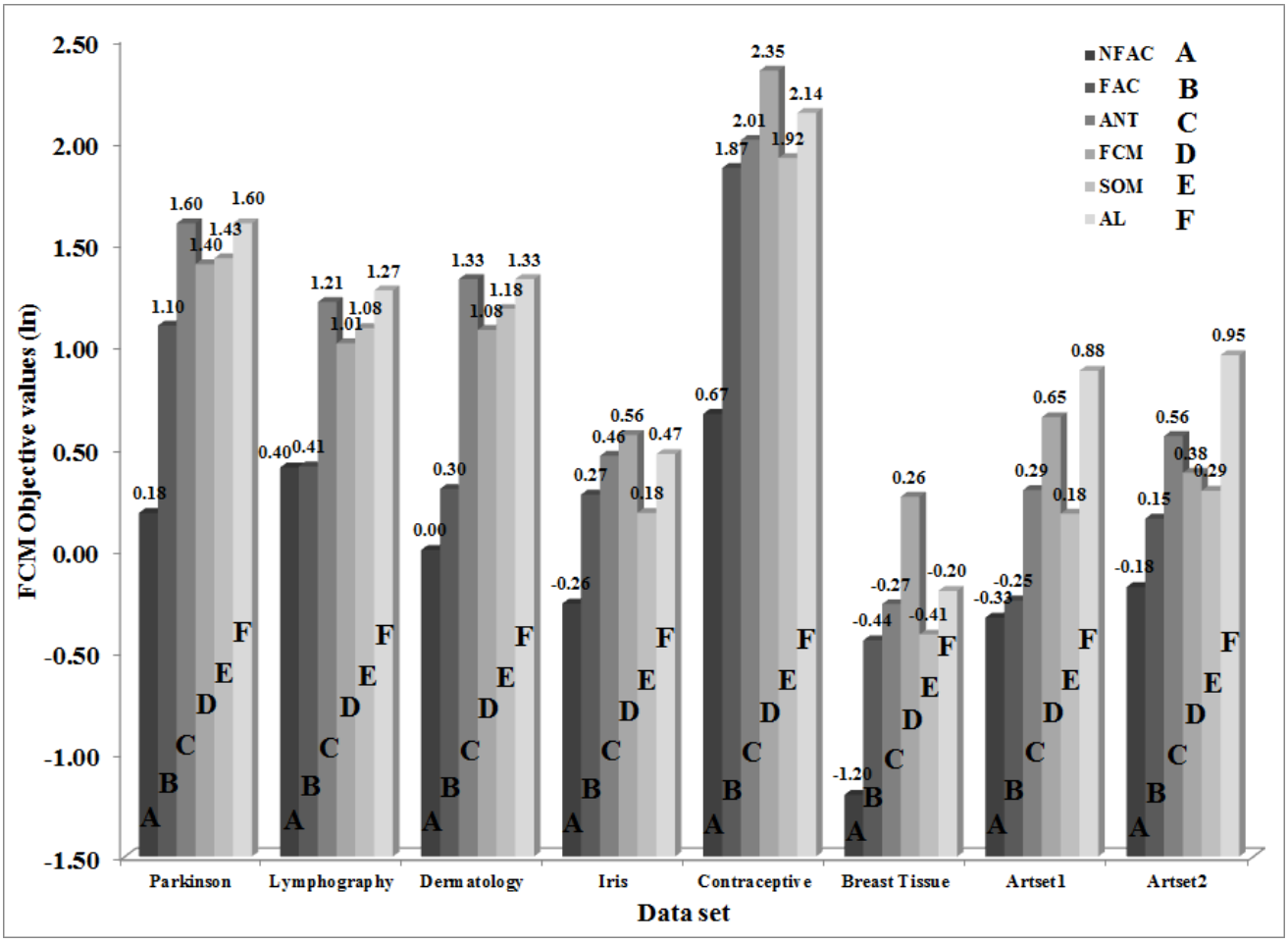


Fig. 8FCM Objective values (ln) generated by each clustering method

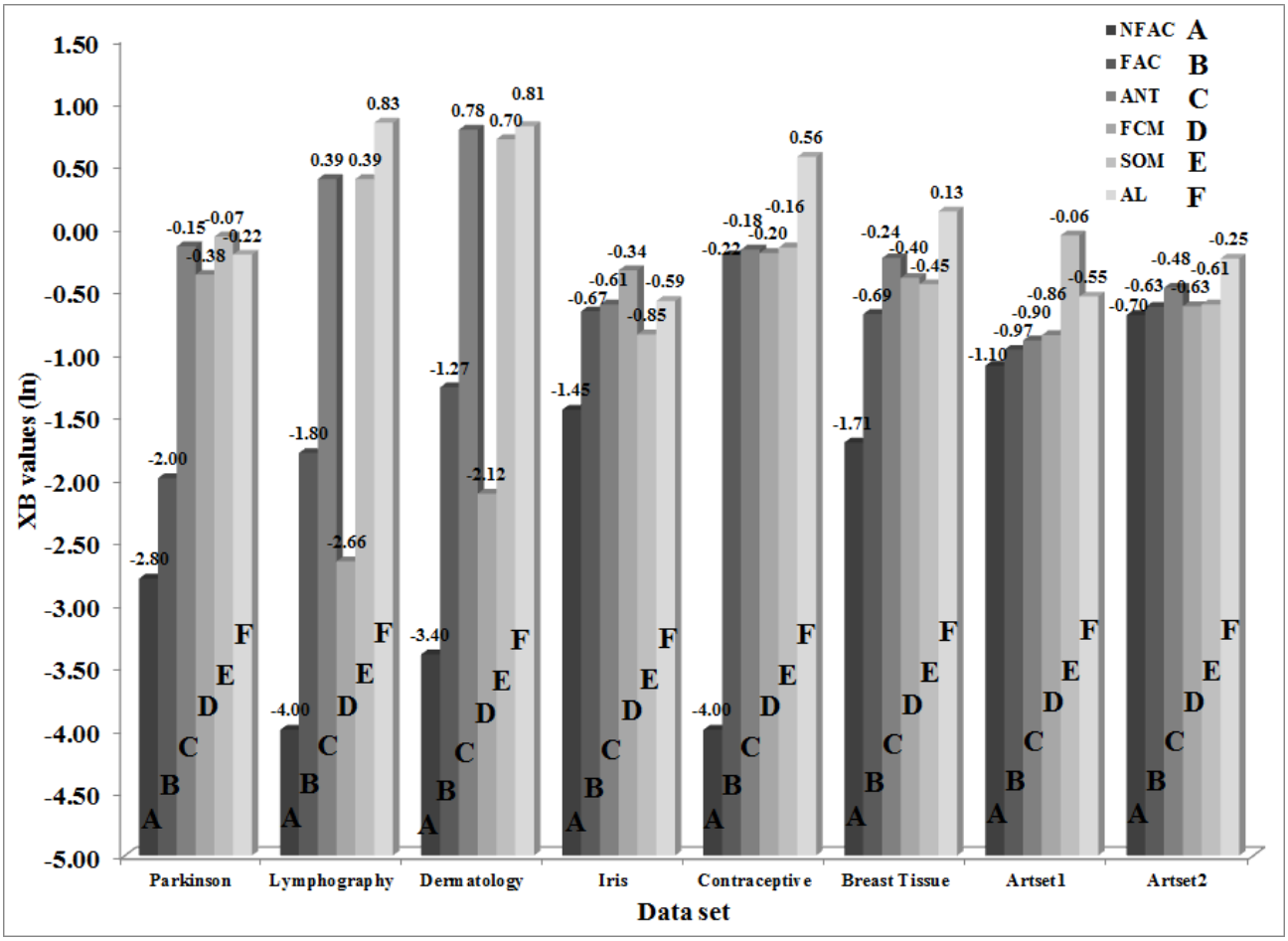
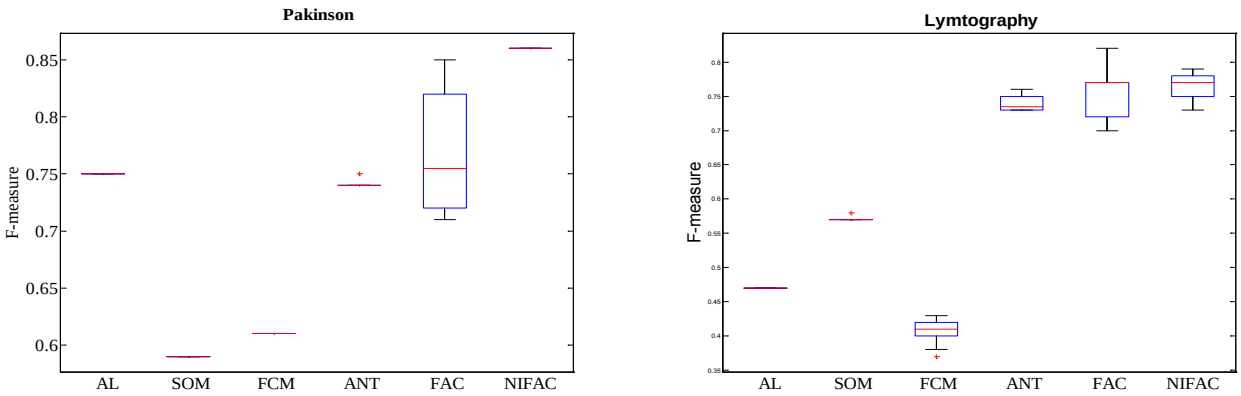


Fig. 9XB values (ln) generated by each clustering method



(a) (b)

Fig. 10 Ranges of F-measure degrees, resulted from running the six clustering algorithms on the eight data sets.

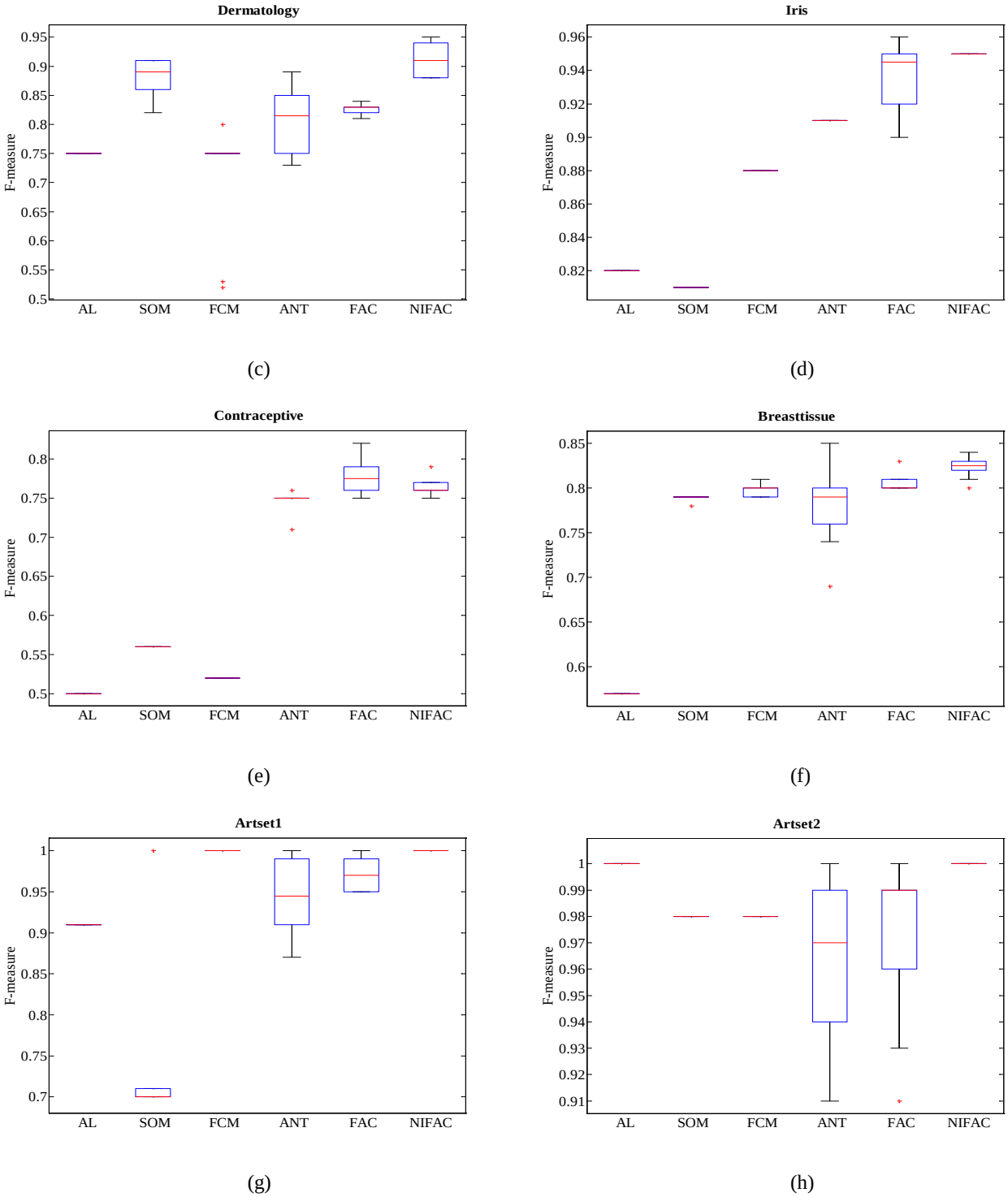


Fig. 10 (Cont.) Ranges of F-measure degrees, resulted from running the six clustering algorithms on the eight data sets.

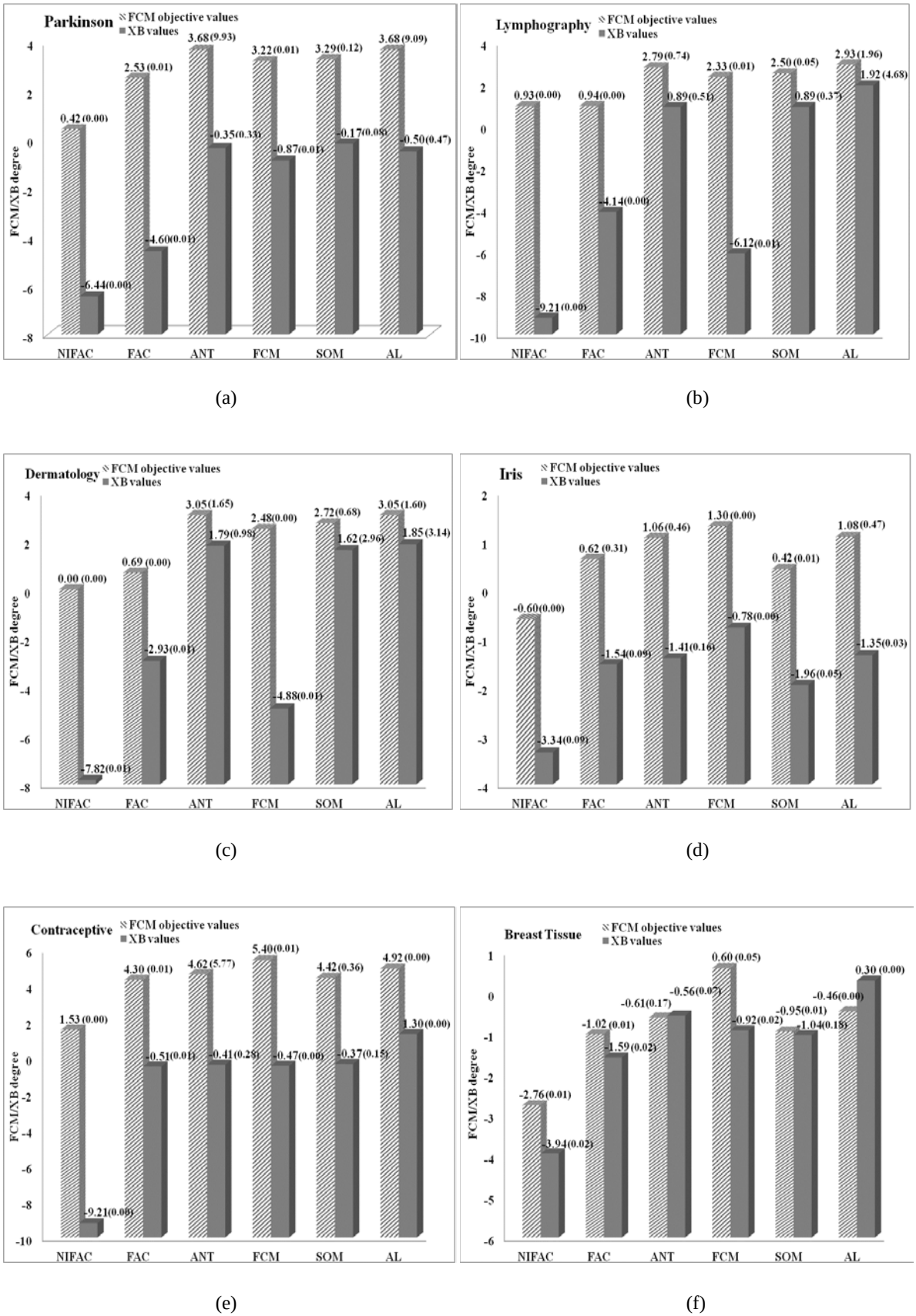
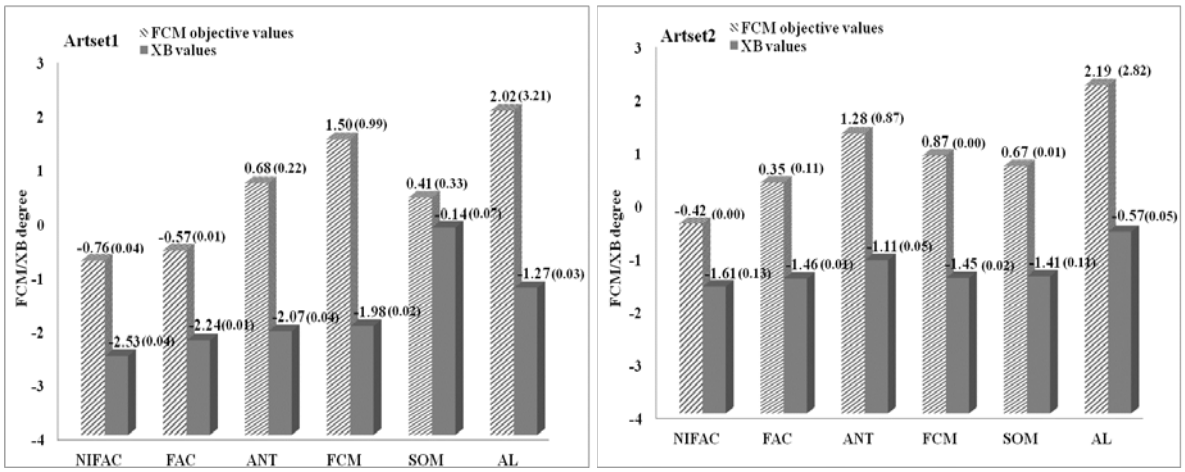


Fig. 11FCM objective degree and XB values resulted from running the six algorithms on the eight data sets. The mean and standard deviations (in parenthesizes) for 10 independent cross-validation runs are reported on the top of the bars.



(g) (h)

Fig. 11 (Cont.) 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.

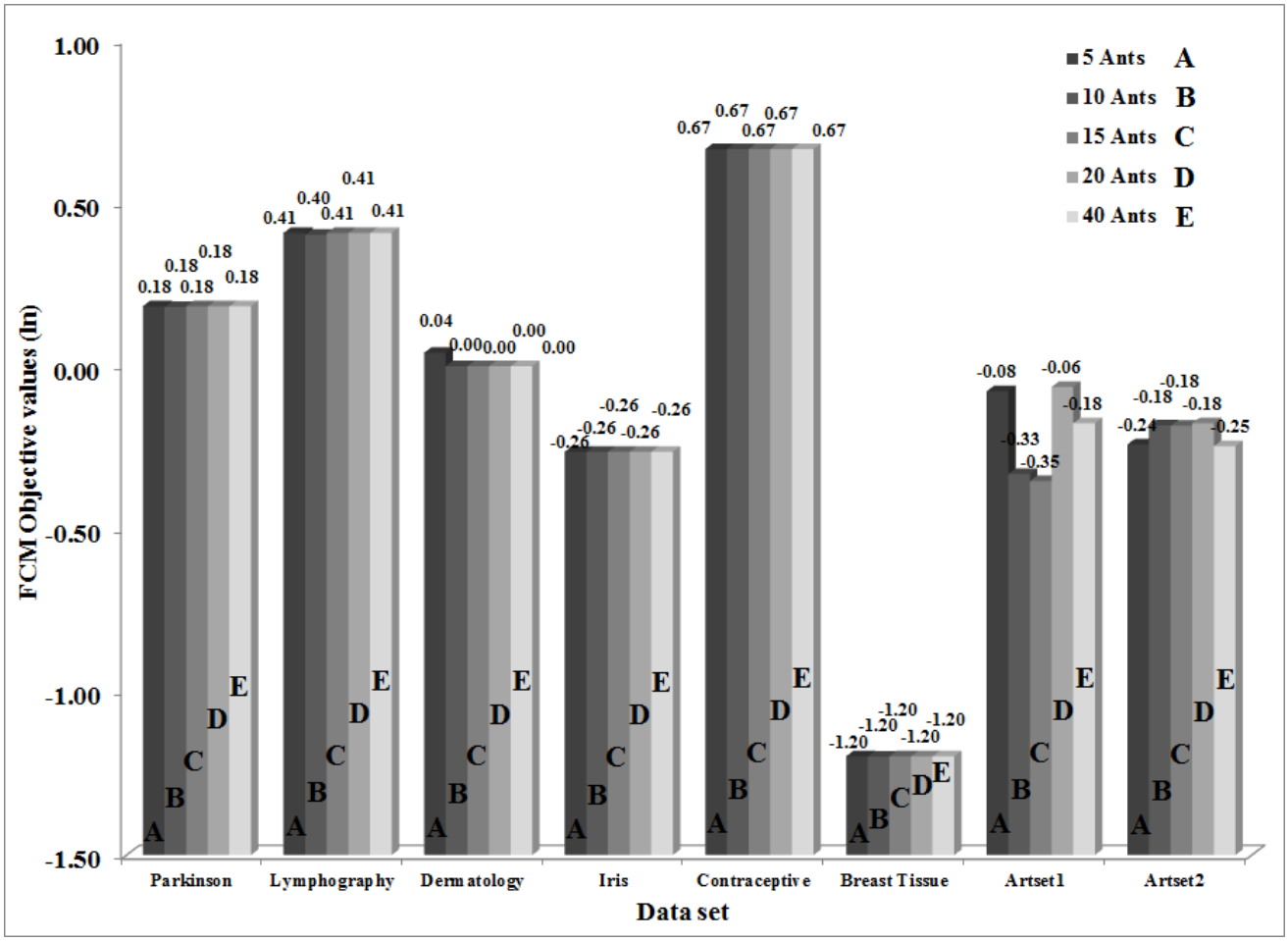


Fig. 12FCM objective values (ln) produced by NIFAC using various numbers of ants

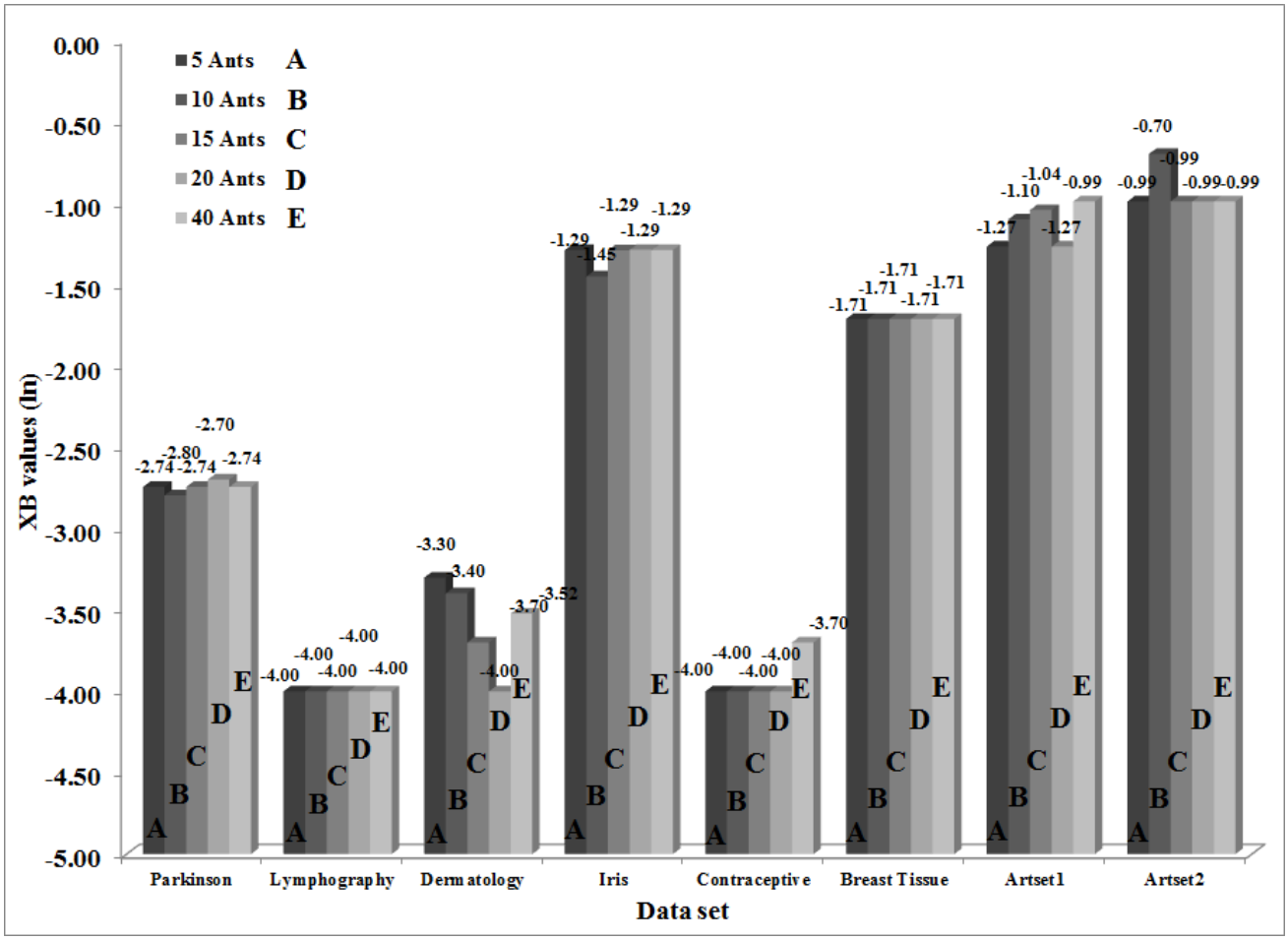


Fig. 13XB values (ln) produced by NIFAC using various numbers of ants

Fig. 11 exhibits the natural logarithmic values of FCM objective and XB degree of the six algorithms. The means and standard deviations (in parenthesizes) of the NIFAC results are remarkable, compared to the others. Fig. 12 and 13 display natural logarithms of FCM objective and XB degree yielded by NIFAC using various numbers of ants. For most cases, all the related methods generate close values of FCM objective degree. The average FCM objective values regarding 5, 10, 15 ,20 and 40 number of ants employed are1.4818, 1.4285, 1.4303, 1.4842 and 1.4473 respectively ; whereas the average XB are 0.0287, 0.0421, 0.0332, 0.0287 and 0.0347. By determining all of these measures, using 10 numbers of ants for NIFAC is enough for generating good results of clustering within the scope of the eight data sets. The most efficient runtime consumption is given by using 5 ants.

VI. Conclusion

This paper presents NIFAC which refers to the novel version of ant-based clustering (ANT) integrating with fuzzy c-means (FCM). The improvement emphasizes on the nonparametric balance of exploitation and exploration mechanisms during ANT search. Striking the balance between exploration and exploitation is one of significant keys to achieve globally best solution. During the iteration runs of NIFAC, all explorative and exploitative steps follow the divide-and-conquer principle, such that a feature space is divided into several different regions that clustering searches are solved independently. Afterwards, optimal partition solutions, yielded by each of those regions seamlessly enter the process of decision making to form completely final partition solutions. In addition, none of arbitrarily defined parameters is employed to control the cycle of exploration and exploitation mechanisms. Such nonparametric mechanisms point the important advantage of NIFAC. The experiment is done on six benchmark real world and two artificial data sets. Among all comparative clustering methods, ANT, FCM, SOM and AL, the proposed NIFAC 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 experiment also reports another merit of NIFAC according to the achievement of the powerful clustering, using a few numbers of ants with a moderate number of runs. However, NIFAC cannot be applicable when the runtime is quite critical. In the future, the methodologies, concerning nonparametric search in some other ways will be focused to achieve better runtime results.

REFERENCES

[1] R. Rojas,*Neural Networks*. Springer-Verlag, Berlin Germany, 1996.

[2] A.Herrero andE.Corchado, "Alfredo J. Unsupervised neural models for country and political risk analysis,"*Expert Systems with Applications*, vol. 38, no. 11, pp. 13641-13661,2011.

[3] A. R.Webb,*Statistical Pattern Recognition*. John Wiley & Sons Ltd., 2002.

[4] P. Tan, M. Steinbach, and V. Kuma,*Introduction to Data Mining*, Addison-Wesley, 2004.

[5] E. Alpaydin,"Introduction to machine learning,"*Cambridge, MA: The MIT Press*, pp. 1-3, 2004.

[6] T. Kohonen,*Self-Organizing Maps*. Springer-Verlag, Berlin Germany, 1995.

[7] T.Hastie,R. Tibshirani, and J. Friedman, "Hierarchical clustering,"*The Elements of Statistical Learning 2nd ed.* New York Springer, pp. 520-528, 2009.

[8] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations,"*In Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, pp. 281-297, 1967.

[9] K.Kerdprasop, S.Taokok, and N.Kerdprasop"Declarative Parallelized Techniques for K-Means Data Clustering,"*International Journal Of Mathematics And Computers In Simulation*, vol. 6, no. 5, pp. 483-495, 2012.

[10] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,"*Journal of Cybernetics*, vol. 3, pp. 32-57, 1973.

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

[12] H. W. Tin, S. W.Leu, and S. H. Chang,"An PSO-based Approach to Speed up the Fractal Encoding,"*International Journal Of Mathematical Models And Methods In Applied Sciences*, vol. 6, no. 3, pp.499-506, 2012.

[13] G.Gan, J.Wu, and Z.Yang, "A genetic fuzzy k-modes algorithm for clustering categorical data,"*Expert Systems with Applications*, vol. 36,pp. 1615-1620, 2009.

[14] H.Izakian, and H.Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem,"*Expert Systems with Applications*, vol. 38, pp. 1835-1838, 2011.

[15] Y. Kao, J. Lin, and S. Huang, "Fuzzy clustering by differential evolution,"*In Proc. Intelligent Systems Design and Applications, ISDA'08.*, pp. 246-250, 2008.

[16] Y. Tian, D. Liu, and H. Qi,"K-harmonic means data clustering with Differential Evolution,"*In Proc. BioMedical Information Engineering (FBIE 2009), International Conference on Future*, pp. 369-372, 2009.

[17] W.Gong, Z.Cai, C. X.Ling, and J. Du,"Hybrid Differential Evolution based on Fuzzy C-means Clustering,"*In proceedings of Genetic and Evolutionary Computation Conference (GECCO 2009),ACM Press*, pp. 523-530, 2009.

[18] S. Supratid, P. Julrode, "A performance comparison using principal component analysis and differential evolution on fuzzy c-means and k-harmonic means,"*Rangsit Journal of Arts and Sciences RJAS*, vol. 1, no. 2,pp. 127-137, 2011.

[19] E.Bonabeau, M.Dorigo, and G. Theraulaz, "Swarm Intelligence: from natural to artificial systems,"*New York, USA: Oxford University Press, Inc.*, 1999.

[20] Z. B. Claude, "Application of the Convergence and Divergence of two functions,"*International Journal Of Mathematical Models And Methods In Applied Sciences*, vol. 6, no. 4, pp. 552-559, 2012.

[21] P. M.Kanade, and L. O.Hall, "Fuzzy ants as a clustering concepts,"*In Proc. of the 22nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, pp. 227-232, 2003.

[22] P. M.Kanade, and L. O.Hall, "Fuzzy Ants and Clustering,"*IEEE Transactions On Systems Man And Cybernetics Part A: Systems And Humans*, vol. 37, no. 5, 2007.

[23] C. R.Bonham, "An investigation of exploration and exploitation within cluster oriented genetic algorithms (COGAs),"*In Proc. of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann*, pp. 1491-1497, 1999.

[24] H. S.Chang, "An ant system based exploration-exploitation for reinforcement learning,"*In Proceedings of the IEEE Conference on Systems, Man, and Cybernetics. IEEE Press, Piscataway, NJ, USA*, pp. 3805-3810, 2004.

[25] C.Fang, J. Lee, and M. A.Schilling, "Balancing Exploration and Exploitation Through Structural Design,"*The Isolation of Subgroups and Organizational Learning Organization Science*, vol. 21, pp. 625-642, 2009.

[26] M. Y. Kiang, "A Comparative Assessment of Classification Methods,"*Decision Support Systems*, vol. 35, no. 4, pp. 441-454, 2003.

[27] D.Li, and C. Yeh, "A non-parametric learning algorithm for small manufacturing data sets,"*Expert Systems with Applications*, vol. 34 no. 1, pp. 391-398, 2008.

[28] R. P.Dinesh, D. L.Kenneth, K. K.Ronald, and M. L.Sheila, "Experimental comparison of parametric, non-parametric, and hybrid multigroup classification," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8593-8603, 2012.

[29] M.Memmedli, and A.Nizamitdinov, "An Application of Various Nonparametric Techniques by Nonparametric Regression Splines," *International Journal Of Mathematical Models And Methods In Applied Sciences*, vol. 6,no. 1, pp. 106-113, 2012.

[30] A. Dursun, "Comparison of regression models based on nonparametric estimation techniques: Prediction of GDP in Turkey,"*International Journal of Mathematical Models And Methods In Applied Sciences*, vol. 1, no. 2, pp.70-75, 2007.

[31] R.Rugina, and M.Rinard, "Recursion unrolling for divide-and-conquer programs. Languages and Compilers for Parallel Computing,"*Lecture Notes in Computer Science (Berlin: Springer, 2001)*, pp. 34-48, 2001.

[32] A. Dalli, "Adaptation of the F-measure to cluster-based Lexicon quality evaluation,"*In EACL 2003*, Budapest, 2003.

[33] J.Handl, J.Knowles, and M. Dorigo, "On the performance of ant-based clustering. Design and application of hybrid intelligent systems,"*Frontiers in Artificial Intelligence and Applications*, vol. 104, pp. 204-213, 2003.

[34] X. L.Xie, and G. Beni, "A validity measure for fuzzy clustering,"*IEEE Trans. Patter Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841-847, 1991.

[35] C. Olson, "Parallel algorithms for hierarchical clustering,"*Parallel Computing*, vol. 21, no. 8, pp. 1313-1325, 1995.

Phichete Julrode received the B.Sc. in computer science from PhetburiRajabhat University, Thailand in 1995. He received his master's degree in computer science from the Chiangmai University, Thailand in 2005. He is currently a lecturer in the informatics department, faculty of science and technology at Phuket Rajabhat University, Phuket, Thailand. His research interests include artificial intelligence, data mining, machine learning, and evolutionary algorithms.

SiripornSupratid is an assistant professor at the Department of Information Technology, Rangsit University, Thailand. She received her bachelor degree in Statistics (Mathematical Statistics) from Chulalongkorn University, Thailand, in 1985, master degree in Management Information Systems (MIS), University of Colorado, USA, 1989 and doctoral degree in computer science from Asian Institute of Technology, Thailand, in 2005. Her research interests include data mining, machine learning, swarm intelligence and fuzzy rule-based.