Rangsit Journal of Arts and Sciences, July-December 2011 RJAS Vol. 1 No. 2, pp. 127-137

A performance comparison using principal component analysis and differential evolution on fuzzy c-means and k-harmonic means

Siriporn Supratid[1]* and Phichete Julrode[2]

Faculty of Information Technology, Rangsit University, Patumthani 12000, Thailand
[1]E-mail: siri_sup1@hotmail.com; [2] E-mail: phichete@pkru.ac.th

*Corresponding author

Submitted 6 August 2011; accepted in final form 27 November 2011

Abstract

Several clustering researches have attempted to optimize the clustering approaches regarding initial clusters. The purpose is to alleviate local optima traps. However, such an optimization may possibly not significantly improve the accuracy rate; contrarily it usually generates abundant runtime consumption. In addition, it may cause the emergence of local traps rather than providing the proper clusters initialization. One may turn to focus on the problems of high dimensional, noisy data and outliers hidden in real-world data. Such difficulties can seriously spoil the computation of several types of learning, including clustering. Feature reduction is one of the approaches to relieve such problems. Thereby, this paper proposes a performance comparison using principal component analysis (PCA) and differential evolution (DE) on fuzzy clustering. The purpose relates to evaluating the consequences of feature reduction, compared to those of optimization of the clustering environment. Here, the fuzzy clustering approaches, fuzzy c-means (FCM) and k-harmonic means (KHM) are experimented. FCM and KHM are soft clustering algorithms that retain more information from the original data than those of crisp or hard. PCA, the feature reduction method, is employed as a preprocessing of FCM and KHM for relieving the curse of high-dimensional, noisy data. The performance of the FCM and KHM based on PCA feature extraction, called PCAFCM and PCAKHM are compared with related algorithms, including the FCM and KHM optimized by differential evolution (DE) method. Comparison tests are performed related to 7 well-known benchmark real-world data sets. Within the scope of this study, the superiority of the feature reduction using PCA over DE optimization on FCM and KHM is indicated.

Keywords*: principal component analysis, fuzzy c-means, k-harmonic means, differential evolution.*

1. Introduction

Many real-world applications are really ambiguous and cannot be exclusively clustered into distinct classes. To compensate for these artifacts, soft clustering methods, fuzzy c-means (FCM) algorithm (Bezdek, Ehrlich, & Full, 1984), has recently been used extensively with some success in the fuzzy clustering areas (Balafar, Ramli, Saripan, Mahmud, & Mashohor, 2008; Chen, Ginger, & Bick, 2006; Chuang, Tzeng, Chen, Wu, & Chen, 2006; Kannan, Ramathilagam, Sathya, & Pandiyarajan, 2010; Yong, Chongxun & Pan, 2004;). The objective regarding FCM is to group data, based on arithmetic means into set of disjointed clusters. The data within the same clusters are highly similar with one another and dissimilar with those in other clusters. The strength of FCM over traditional clustering such as k-means (KM) is that it allows one piece of data to belong to two or more clusters. Given an input point, FCM yields the degree of membership value in each cluster. The other presented clustering method, k-harmonic means (KHM) algorithm is an algorithm proposed by Zhang, Hsu, and Dayal (1999), Zhang (2000) and modified by Hammerly and Elkan (2002). Instead of using an arithmetic mean, k-harmonic clustering uses the harmonic average to calculate the similarity of the points in the data set to the cluster centers, to which they belong. Moreover, KHM associates the influence weight of a single data on the cluster center in the following iterations. FCM as well as KHM are soft clustering algorithms that retain more information from the original data than those of crisp or hard; they perform data clustering based on arithmetic and harmonic means. KHM also associates the influence weight of a single data on the cluster center for the next iteration (Frackiewicz & Palus, 2008; Li, Gu, & Zhang, 2010; Ma & Staunton, 2007).

There still exists a question on the causes of local optima traps problems, occurring in FCM

as well as KHM learning. In order to avoid such local traps, many researches applied several types Thangaraj, 2010; Wang, Liu, Zhao, & Xu, 2006; Yang, Sun & Zhang, 2009). Differential evolution (DE), one of the efficient optimization methods, has been introduced by (Price, Storn, & Lampinen, 2005; Storn & Price, 1997). The main idea is to create a population of candidate solutions to an optimization problem. Those solutions are iteratively refined by alteration and selection of good solutions for the next iteration. FCM and KHM are sensitive to the initial clusters. There exists research involved with optimizing FCM or KHM by using DE. FCDE, an algorithm of FCM optimized by DE, was proposed by (Kao, Lin, & Huang, 2008). FCDE showed better performance over FCM. The work (Tian, Liu, & Qi, 2009) presented an algorithm concerning KHM, optimized by DE for the clustering purpose. Such an algorithm took an advantage of DE's global searching ability to overcome getting stuck at local minima. The results clearly demonstrated that the optimization algorithm obtained more acceptable results than KHM.

Nevertheless, such clustering optimizations may not significantly improve clustering precision; oppositely, the runtime consumption is usually worse than the clustering without any optimization. This is a reason why one should turn to focus on some other cause of local trap problems such as high dimensional, noisy data and outliers. Such problems seriously spoil the computation of several types of learning, including clustering. Several real-world applications usually suffer from such high dimensionality problems. Irrelevant dimensional features could seriously deteriorate the generalization performance of clustering. A linear feature extraction method, principal component analysis (PCA) is one of the important tools for coping with such dimensionality problems (Jolliffe, 1986). In order to perform dimension reduction, PCA maps the original predicting features into smaller numbers of features. Thereby, applying PCA as preprocessing for such a dimension reduction would lead to the improvement of the clustering efficiency. There have been some works e.g., the library evaluation (Wei & Li, 2009) that combined PCA with FCM. Such a work denoted the advantage of a cooperation of PCA and FCM over FCM and the Back Propagation neural network. of optimizations on either FCM or KHM to get the appropriate initial set of clusters (Gomathi & This work gave the incentive to this study. The optimization and feature reduction, related to clustering areas, have been separately explored. A comparison study among them has not yet been fulfilled.

This paper presents a performance comparison using principal component analysis (PCA) and using differential evolution (DE) on fuzzy clustering. The combination of the preprocessing and the clustering approaches are here called, PCAFCM and PCAKHM. The optimizations of DE on FCM as well as on KHM, called here DEFCM and DEKHM are also determined in a comparison test. Here, PCAFCM and PCAKHM are compared to DEFCM and DEKHM. The rest of the paper is organized as follows. Sections 2 and 3 introduce FCM and KHM clustering. Section 4 briefly describes DE search technique. In section 5, PCA preprocessing is described. Then, experimental results are determined in section 6. Finally, conclusions are made in section 7.

2. Fuzzy c-means (FCM)

Fuzzy C-Means (FCM) is a clustering method that allows a data point to belong to two or more clusters with different degrees of membership; unlike k-means (KM), the traditional clustering method that assigns a pattern to only a single cluster. FCM is widely used in pattern recognition. It is based on minimization of the following objective function:

$$\sum_{n=1}^{N}\sum_{c=1}^{C} u_{nc}^{m}\|x_n - \bar{x}_c\|^2 \qquad (1)$$

where, $\mathbf{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n, ..., \boldsymbol{x}_N\}$, $\boldsymbol{x}_n$ is the n^{th} of d-dimensional measured data; is a set of data to be clustered $\boldsymbol{x}_c$ is a c^{th} cluster centers, where $c = 1, 2, \ldots, C$. m, fuzziness degree controls the extent of membership sharing between fuzzy clusters; here it equals 2, $\boldsymbol{u}_{nc}$ is the degree of membership of input $\boldsymbol{x}_n$ in the cluster c. $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in (1). The update of membership $\boldsymbol{u}_{nj}$ and the cluster centers $\boldsymbol{x}_c$ follow (2) and (3) consecutively:

$$u_{nj} = \left(\sum_{c=0}^{n} \frac{\left\| x_n - \bar{x}_j \right\|^{\frac{2}{m-1}}}{\| x_n - \bar{x}_c \|} \right)^{-1} \quad (2)$$

$$\bar{\boldsymbol{x}}_j = \frac{\sum_{c=1}^{C} \boldsymbol{u}_{nj}^{m} \boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{u}_{nj}} \quad (3)$$

This iteration will stop when:

$$max\left\{ \left| \boldsymbol{u}_{nc}^{iter+1} - \boldsymbol{u}_{nc}^{iter} \right| \right\} < \varepsilon$$

where, ε is a termination criterion ranged between 0 and 1 and superscript *iter* is the iteration number. However, the problem of getting into local optima still exists in FCM learning.

3. K-harmonic means clustering (KHM)

K-harmonic means clustering (KHM) applies degrees of membership to allow each data point to belong to two or more clusters (Hammerly & Elkan, 2002; Gungor & Unler, 2007; Gungor & Unler, 2008), similar to FCM. However, in KHM, the arithmetic mean of distance from a data point to the centers, used in FCM, is replaced by the harmonic mean. The harmonic means gives a good (low) score for each data point when that data point is close to any one center. This is a property of the harmonic means; it is similar to the minimum function used by KM, but it is a smooth differentiable function. The following notations are used to formulate the KHM algorithm.

$p(c/\boldsymbol{x}_n)$: the membership function defining theproportion of data point $\boldsymbol{x_n}$ that belongs to center *c*.

$w(\boldsymbol{x}_n)$: the weight function defining how much influence data point $\boldsymbol{x_n}$ has in re-computing the center parameters in the next iteration.

The basic algorithm for KHM clustering is shown as follows:

1. Randomly choose the initial *C* centers.
2. Calculate the objective function value according to

$$KHM(\mathbf{X}, C) = \sum_{n=1}^{N} \frac{C}{\sum_{c=1}^{C} \frac{1}{\| \boldsymbol{x}_n - \bar{\boldsymbol{x}}_c \|^m}} \quad (4)$$

where *m* is equivalent to 2, like that in FCM.

3. For each data point $\boldsymbol{x}_n$, compute its member-ship $p(c/\boldsymbol{x}_n)$ in each center $\bar{\boldsymbol{x}}_c$ according to (5)

$$p\left(\frac{c}{\boldsymbol{x}_n}\right) = \frac{\| \boldsymbol{x}_n - \bar{\boldsymbol{x}}_c \|^{-m-2}}{\sum_{c=1}^{C} \| \boldsymbol{x}_n - \bar{\boldsymbol{x}}_c \|^{-m-2}} \quad (5)$$

4. For each data point $\boldsymbol{x}_n$, compute its weight $w(\boldsymbol{x}_n)$ according to (6)

$$w(\boldsymbol{x}_n) = \frac{\sum_{c=1}^{C} \| \boldsymbol{x}_n - \bar{\boldsymbol{x}}_c \|^{-m-2}}{(\sum_{c=1}^{C} \| \boldsymbol{x}_n - \bar{\boldsymbol{x}}_c \|^{-m})^2} \quad (6)$$

5. For each center $\bar{\boldsymbol{x}}_c$, re-compute its location from all data points $\boldsymbol{x_n}$ according to their member-ships and weights as show in (7)

$$\bar{\boldsymbol{x}}_c = \frac{\sum_{n=1}^{N} p(c/\boldsymbol{x}_n) w(\boldsymbol{x_n}) \boldsymbol{x_n}}{\sum_{i=1}^{n} p(c/\boldsymbol{x}_n) w(\boldsymbol{x_n})} \quad (7)$$

6. Repeat steps 2–5 until reaching a predefined number of iterations or until KHM(**X**,*C*) does not change significantly.
7. Assign a data point $\boldsymbol{x_n}$ to a cluster $\bar{\boldsymbol{x}}_c$ with the biggest $p(c/\boldsymbol{x}_n)$.

4. Using differential evolution for optimizing FCM and KHM (DEFCM and DEKHM)

Differential evolution (DE) algorithm (Price, Storn & Lampinen, 2005) is a simple and efficient heuristic for global optimization over continuous spaces. Like any other evolutionary algorithm, DE is also a stochastic population-based method. However, DE does not make use of some probability distribution function in order to introduce variations into the population. Instead, DE uses the differences between randomly selected data vectors as the source of random variations for a third vector, referred to as the target vector. Trial solutions are generated by adding weighted difference vectors to the target vector. This process is referred to as the mutation operator where the target vector is mutated. The crossover and selection step are then applied to produce an offspring which is only accepted if it improves on the fitness of the parent individual.

DE is used for initial clusters optimization on FCM and KHM. The optimization, made on FCM and KHM are consecutively called DEFCM

and DEKHM. The idea is similar to other clustering optimization such as simulated annealing heuristic (SA) (Tian, Liu, & Qi, 2009) or tabu search method (Bankapalli, Babu, & Devi, 2011). In DEFCM and DEKHM, the populations of candidate solutions which contain *C* number of cluster centers are sampled randomly. Then the standard KHM and DE algorithm executes alternately when the stop criterion is satisfied, and the final solution of the cluster centers and the partition obtained by the globally best chromosome is reported. The pseudo code for the complete DEFCM and DEKHM clustering algorithm is presented in Figure 1.

5. Using principal component analysis with FCM and KHM for dimension reduction (PCAFCM and PCAKHM)

PCA is an orthogonal basic transformation. Given a data set: $\{\boldsymbol{x}_i \in \boldsymbol{R}^D | i = i, \dots, N\}$, where *D* is the number of dimensions, *N* refers to the samples size. $\mathbf{Y} = (y_1,\dots,y_n)$ is given as a centered matrix; $y_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}$, where $\overline{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_i / N$. The basis is found by diagonalizing the centered covariance matrix, defined by

$$\mathbf{M} = \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\boldsymbol{T}} = \mathbf{Y}\mathbf{Y}^{\mathbf{T}}$$

The coordinates in the eigenvector basis are called *principal components*. In PCA, one has to find eigenvalues λ_p and eigenvectors $\boldsymbol{v}_p$ of **M**, satisfying $\boldsymbol{v}_p = \mathbf{M}v_p$. The size of each eigenvalue λ_p equals the amount of variance in the direction of the corresponding an eigenvectors $\boldsymbol{v_p}$where $p = 1,2, \dots, P, P \leq D$. The directions of the first eigenvectors corresponding to the biggest eigenvalues cover as much variance as possible by *P* orthogonal directions. The principal eigenvectors $\boldsymbol{v_p}$ of **M** are the principal directions of $(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$. The principal eigenvectors $\boldsymbol{u}_p$ of **M** are the principal components. Entries of each $\boldsymbol{u}_p$ are the projected values of data points on the principal direction $\boldsymbol{u}_p$. $\boldsymbol{v}_p$ and $\boldsymbol{v}_p$ are related via (8)

$$\boldsymbol{u_p} = \mathbf{Y}^{\mathbf{T}}\boldsymbol{v_p} / \sqrt{\lambda_p} \qquad (8)$$

DEFCM and DEKHM algorithms

Step 1. Initialize each candidate solutions to contain *C* number of randomly selected cluster centers (population size is *NumPop* and dimension is *D*)
Step 2. Set iterative count *t=0* and Maximum number of iterations *maxIter.*
Step 3. Execute DE algorithm on population.
  Step 3.1. *(mutation)* Generate a mutant vector.
  Step 3.2. *(crossover)* Generate a trial vector.
  Step 3.3. Compare trial vector and target vector by their fitness, and update the globally best candidate solution *G*.
Step 4. For each candidate solutions, execute basic FCM (or KHM)
  Step 4.1. Compute membership function and weight value, then update cluster centers.
  Step 4.2. Calculate the objective function value following (1) and (4) and update the globally best candidate solution *G.*
Step 5. If *t* < *maxIter*, set *t* = *t* + *1*, and goto Step 2; otherwise, output the final solution with cluster centers and the partition obtained by the globally best candidate solution *G.*

Figure 1 The pseudo-code of DEFCM and DEKHM.

Through such a PCA method, the main *P* of *D* dimensions are extracted; whilst noisy and irrelevant dimensional features, that could seriously deteriorate the generalization performance of clustering are eliminated. Thereby, PCA is utilized in this paper as preprocessing and generates data with dimension reduction for later used in FCM and KHM learning process. The combination of PCA pre-processing and the clustering approaches are here called, PCAFCM and PCAKHM.

6. Experimental and results

PCAFCM and PCAKHM are tested on seven benchmark medical data sets used from the URL, "http://archive.ics.uci.edu/ml/datasets.html" (Frank & Asuncion, 2010), PCAFCM as well as PCAKHM also compared with FCM, DEFCM, KHM and DEKHM. According to PCA feature reduction, the number of orthonormal eigenvectors, corresponding to the first 90% largest eigenvalues of the covariance matrix would be used in the feature space. Such a selection

criterion of eigenvector is applied for all data sets and related methods. This provides fairness for comparison determination. The details of the tested data sets are described in section 6.1.

6.1 Data sets

6.1.1 Pima Indians Diabetes (n = 768, d = 8, k = 2), which consists of 768 objects characterized by eight features: number of times pregnant, plasma glucose concentration at 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and age. There are two categories in the data: tested positive for diabetes (268 objects) and vice versa (500 objects).

6.1.2 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 data: Parkinson's (147 objects) and healthy (48 objects).

6.1.3 Lymphography (n = 148, d = 18, k = 4), which consists of four different types of lym: normal find (2 objects), metastases (81 objects), malign lymph (61 objects), and fibrosis (4 objects). Each type has eighteen features, which are lymphatics, block of affere, block of lymph. c, block of lymph. s, by pass, extravasates, regeneration, early uptake, lym.nodes dimin, lym.nodes enlar, changes in lym., defect in node, changes in node, changes in stru, special forms, dislocation, exclusion and no. of nodes.

6.1.4 Hepatitis (n = 155, d = 19, k = 2), which consists of two different types of life: dead (32 objects), live (123 objects). Each type has nineteen features, which are age, sex, steroid, antiviral, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime and histology.

6.1.5 Dermatology (n = 366, d = 34, k = 6), which consists of 366 objects 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, munro microabcess, focal hypergranulosis, disappearance of the granular layer, vacuolisation and damage of basal layer, spongiosis, saw-tooth appearance of retes, follicular horn plug, perifollicular parakeratosis, inflammatory monoluclear infiltrate and band-like infiltrate. There are six categories in the data: psoriasis (112 objects), seboreic dermatitis (61 objects), lichen planus (72 objects), pityriasis rosea (49 objects), cronic derma- titis (52 objects) and pityriasis rubra pilaris (20 objects).

6.1.6 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 objects, long-term methods have 334 objects, and short-term methods have 510 objects) of a woman based on her demographic and socioeconomic characteristics.

6.1.7 Breast Tissue (n = 106, d = 9, k = 6): These data, consisting of 106 objects 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: carcinoma (21 objects), fibro-adenoma (15 objects), mastopathy (18 objects), glandular (16 objects), connective (14 objects) and adipose (22 objects).

The characteristics of such medical data sets are summarized in Table 1.

6.2 Experimental results

The experimental results are averages of 10 cross-validation runs. The algorithms are implemented using MATLAB R2010a and executed on an Intel (R) Core(TM)2 Quad CPU 2.40 GHz with 4.00 GB RAM. The quality of the respective clustering are compared, where the quality is measured by the following two criteria:

6.2.1 The objective functions FCM and KHM: the sum over all data points of respectively arithmetic and harmonic average of the distance from a data point to all the centers, as defined consecutively in Eq. (1) and (4). Clearly, the smaller the sum is, the higher the quality of clustering.

6.2.2 Adjusted rand index (ARI): suppose T is the true clustering of a data set based on domain knowledge and R a clustering result given by some clustering algorithm. Let a, b, c, and d, respectively, denote the number of pairs belonging to the same cluster in both T and R, the number of pairs belonging to the same cluster in T but to different clusters in R, the number of pairs belonging to differrent clusters in T but to the same cluster in R and the number of pairs belonging to different clusters in both T and R. The ARI(T,R) is then defined as follows:

$$ART(T,R) = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \quad (9)$$

The value of ARI(T,R) lies between zero and one and higher value indicates that R is more similar to T. In addition, ARI(T,T) = 1.

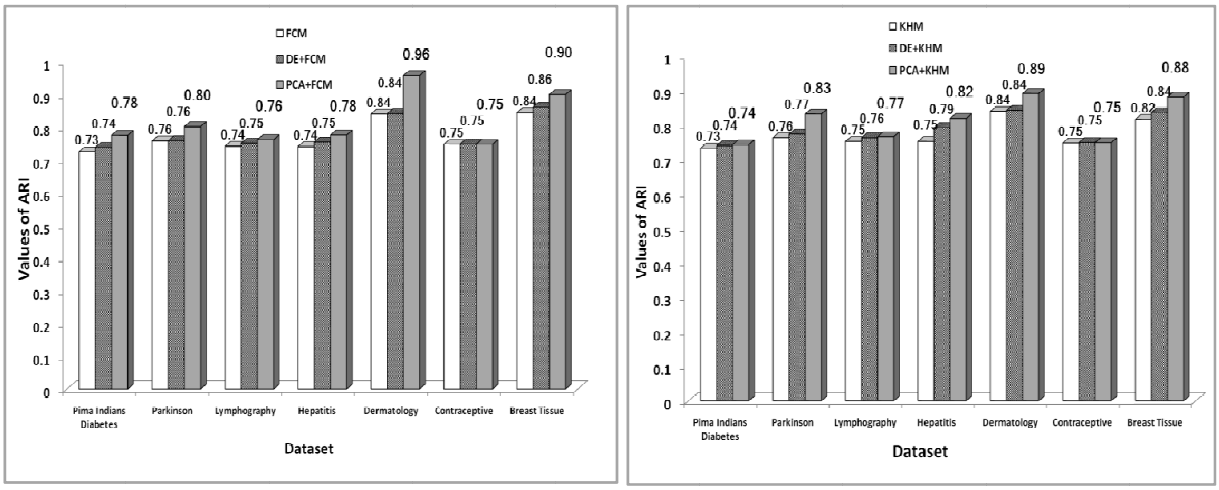
Table 1 Characteristics of data sets considered

Name of data set	No. of classes	No. of features	Size of data set (size of classes in parentheses)
Pima Indians Diabetes	2	8	768 (500, 268)
Parkinson	2	22	195 (48, 147)
Lymphography	4	18	148 (2, 67, 46, 33)
Hepatitis	2	19	155 (32, 123)
Dermatology	6	34	366 (112, 61, 72, 49, 52, 20)
Contraceptive	3	9	1473 (629, 333, 511)
Breast Tissue	6	9	106 (21, 15, 18, 16, 14, 22)

6.2.3 Runtimes: the runtimes of the related clustering algorithms is shown in seconds.

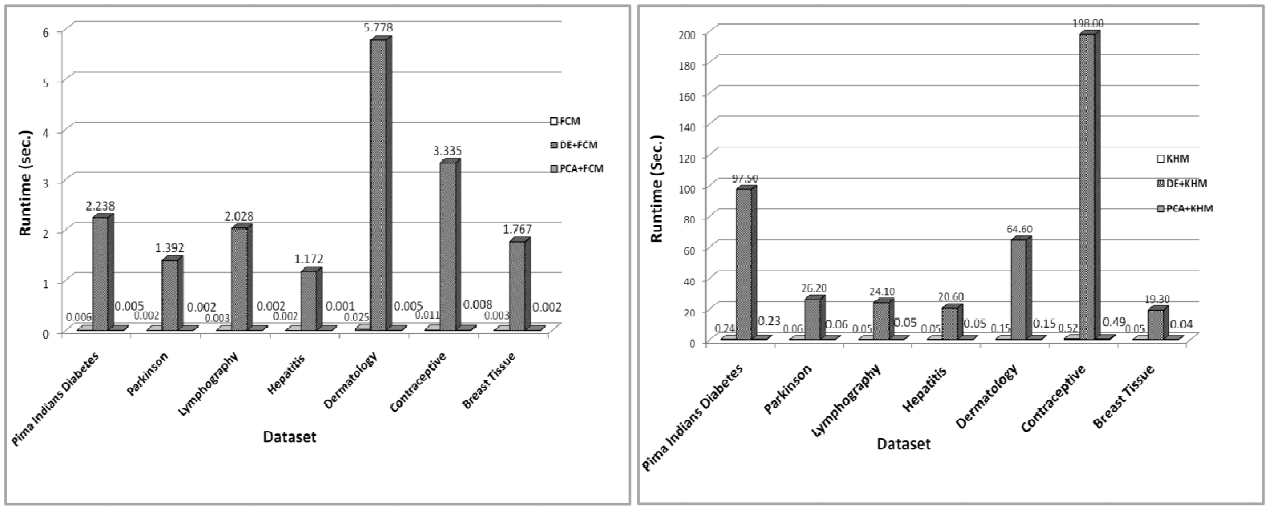
All of these measurement criteria are com- puted as means and standard deviations over 10 independent runs. The results in terms of FCM/KHM objective functions, ARI and runtime are shown in Table 2 and Table 3. Such results indicate most of superiority of PCAFCM and PCAKHM over the others related. A visual analysis of ARI and runtime comparisons can be achieved by determining them in a form of bar graphs. Figure 2 provides a complement to the visual analysis.

In addition, in Table 4 and Table 5, the increase of ARI in percentage, as well as the ratio of runtime based on FCM and KHM are compared between a pair of DEFCM, PCAFCM and that of DEKHM, PCAKHM. The comparison shows the unrivaled improvement of ARI and runtime, yielded by PCAFCM and PCAKHM. The prominent decrease of runtime, belonging to PCAFCM and PCAKHM is explicitly caused by the dimension reduction of the data sets. Contrarily, an abundant multiplication of runtimes is produced by DEFCM and DEKHM. In Figure 3, the dimension reduction is described in a form of bar graph. The stripped and gray bars respectively represent the original number of data dimensions and those of the reduced. The percentages of dimension reduction for each data set are pointed above the gray bars as well.



(a) **(b)**

Figure 2 The measurement (a) ARI value of FCM, DEFCM and PCAFCM, (b) ARI value of KHM, DEKHM and PCAKHM.



(c) **(d)**

Figure 2 (Continued) The measurement (c) Runtime of FCM,DEFCM and PCAFCM, and (d) Runtime of KHM, DEKHM and PCAKHM.

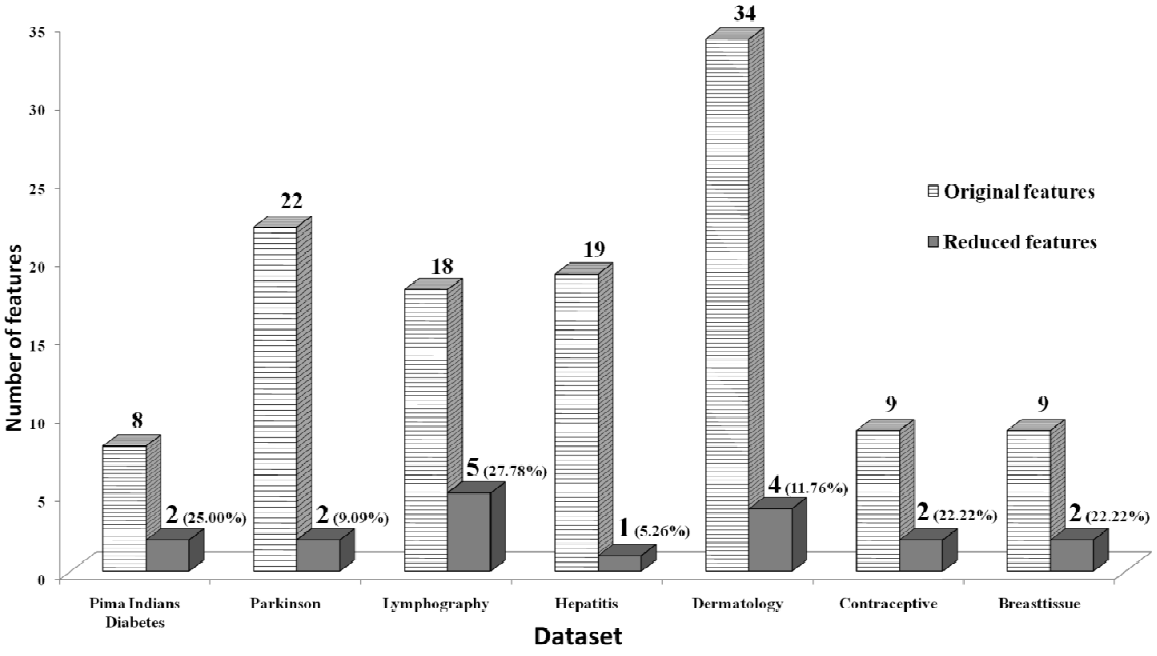


Figure 3 The dimension reduction on seven medical data sets, yielded from PCA as preprocessing for FCM and KHM. The stripped and gray bars represent the original number of data dimensions and those of the reduced. The percentages of dimension reduction for each data set are shown on the gray bars.

Table 2 Results of KHM, DEKHM, and PCAKHM clustering on seven medical data sets when $m = 2$. The quality of clustering is evaluated using KHM objective function and ARI. Runtimes (seconds) are additionally provided. The table shows means and standard deviations (in brackets) for 10 independent cross-validation runs. The subscript [*a* x *b*], attached with each data sets equivalents to [(*size of the data set*) x (*no. of original dimensions*)]. Bold face indicates the best result out of the three algorithms.

Source	KHM	DEKHM	PCAKHM
Pima Indians Diabetes *[8x768]*			
KHM	8.24E+06(3.79E+05)	8.82E+06(4.03E+05)	7.28E+01 (4.45E+00)
ARI	0.732220(0.006228)	0.739340(0.005492)	0.742667(0.009799)
Runtime	0.241481(0.002027)	9.75E+01(3.366258)	0.234806(0.008729)
Parkinson *[22x195]*			
KHM	1.94E+06(7.11E+04)	2.10E+06(7.80E+04)	1.79E+01(3.63E+00)
ARI	0.762732(0.009739)	0.772957(0.010167)	0.831224(0.039765)
Runtime	0.061537(0.000327)	2.62E+01(0.304307)	0.059834(0.000143)
Lymphography *[18x148]*			
KHM	1.85E+03(1.55E+01)	1.87E+03(1.67E+01)	5.71E+01(3.55E-02)
ARI	0.751757(0.007302)	0.762347(0.007152)	0.765476(0.007262)
Runtime	0.053220(0.000383)	2.41E+01(0.105682)	0.052691(0.003550)
Hepatitis *[19x155]*			
KHM	2.03E+06(1.04E+05)	2.15E+06(1.15E+05)	7.40E+00(0.00E+00)
ARI	0.751643(0.024252)	0.792138(0.024214)	0.818666(0.000000)
Runtime	0.049371(0.000213)	2.06E+01(0.091276)	0.053893(0.004582)
Dermatology *[34x366]*			
KHM	4.35E+04(1.41E+03)	4.37E+04(1.48E+03)	9.17E+01(1.67E+00)
ARI	0.839052(0.002650)	0.841695(0.002639)	0.890813(0.003983)
Runtime	0.151916(0.001581)	6.46E+01(0.065226)	0.145047(0.003396)
Contraceptive *[9x1473]*			
KHM	3.35E+04(1.62E+03)	5.39E+04(1.66E+03)	1.46E+02(1.48E+01)
ARI	0.747069(0.001669)	0.748767(0.001589)	0.748210(0.006695)
Runtime	0.516299(0.003731)	1.98E+02(0.196091)	0.489418(0.003203)
Breast Tissue *[9x106]*			
KHM	2.73E+10(1.92E+09)	3.09E+10(2.33E+09)	4.12E+00(5.66E-01)
ARI	0.815894(0.041366)	0.836069(0.040231)	0.880558(0.007862)
Runtime	0.048622(0.000405)	1.93E+01(0.103529)	0.042208(0.003567)

Table 3 Results of FCM, DEFCM, and PCAFCM clustering on seven medical data sets when $m = 2$. The quality of clustering is evaluated using FCM objective function and ARI. Runtimes (seconds) are additionally provided. The table shows means and standard deviations (in brackets) for 10 independent cross-validation runs. The subscript [*a* x b], attached with each data sets equivalents to [(*size of the data set*) x (*no. of original dimensions*)]. Bold face indicates the best result out of the three algorithms.

Source	FCM	DEFCM	PCAFCM
Pima Indians Diabetes *[8x768]*			
FCM	1.77E+06(1.19E+05)	1.68E+06(2.89E+04)	2.62E+02(1.38E+00)
ARI	0.725371(0.005467)	0.738265(0.005410)	0.775219(0.018962)
Runtime	0.005733(0.000324)	2.237780(0.012194)	0.004801(0.000645)
Parkinson *[22x195]*			
FCM	3.78E+05(1.13E+04)	3.70E+05(1.65E+02)	6.11E+01(3.16E+00)
ARI	0.758036(0.015851)	0.758806(0.002230)	0.801960(0.016747)
Runtime	0.002048(0.000089)	1.391502(0.003675)	0.001593(0.000034)
Lymphography *[18x148]*			
FCM	6.11E+01(9.94E-02)	6.09E+01(1.56E-02)	3.22E+01(1.52E-03)
ARI	0.741849(0.004615)	0.751416(0.007600)	0.762226(0.003660)
Runtime	0.002878(0.000164)	2.028351(0.009952)	0.002337(0.000111)
Hepatitis *[19x155]*			
FCM	3.83E+05(7.20E+03)	3.76E+05(6.05E+03)	1.89E+01(1.80E+00)
ARI	0.740921(0.013361)	0.754764(0.013297)	0.776110(0.009413)
Runtime	0.001731(0.000019)	1.172003(0.005808)	0.001280(0.004808)

Table 3 (continued)

Source	FCM	DEFCM	PCAFCM
Dermatology [34x366]			
FCM	7.64E+02(1.36E+01)	7.51E+02(1.41E+01)	1.65E+01(1.52E-01)
ARI	0.840308(0.002782)	0.841562(0.002628)	0.958336(0.015957)
Runtime	0.025276(0.037031)	5.778450(0.150970)	0.004810(0.004928)
Contraceptive [9x1473]			
FCM	5.03E+03(1.86E+02)	4.88E+03(1.57E+02)	1.71E+02(2.76E+00)
ARI	0.747943(0.001083)	0.748656(0.000630)	0.747891(0.001881)
Runtime	0.010990(0.001211)	3.334728(0.015674)	0.008319(0.004156)
Breast Tissue [9x106]			
FCM	3.16E+08(1.96E+07)	3.81E+07(1.27E+07)	1.65E+00(4.72E-02)
ARI	0.844279(0.020612)	0.861057(0.013693)	0.900135(0.009301)
Runtime	0.002773(0.000065)	1.767047(0.003042)	0.002344(0.000059)

Table 4 Percentage increase of ARI base on FCM

Source	% increase of ARI based on FCM		Ratio of Runtime based on FCM	
	DEFCM	PCAFCM	DEFCM	PCAFCM
Pima Indians Diabetes[8x768]	1.78	6.87	390.33	0.84
Parkinson[22x195]	0.10	5.79	679.44	0.78
Lymphography[18x148]	1.29	2.75	704.78	0.81
Hepatitis[19x155]	1.87	4.75	677.07	0.74
Dermatology[34x366]	0.15	14.05	228.61	0.19
Contraceptive[9x1473]	0.10	-0.01	303.43	0.76
Breast Tissue[9x106]	1.99	6.62	637.23	0.85

Table 5 Percentage increase of ARI base on KHM

Source	% increase of ARI based on KHM		Ratio of Runtime based on KHM	
	DEKHM	PCAKHM	DEKHM	PCAKHM
Pima Indians Diabetes[8x768]	0.97	1.43	0.97	1.43
Parkinson[22x195]	1.34	8.98	1.34	8.98
Lymphography[18x148]	1.41	1.82	1.41	1.82
Hepatitis[19x155]	5.39	8.92	5.39	8.92
Dermatology[34x366]	0.31	6.17	0.31	6.17
Contraceptive[9x1473]	0.23	0.15	0.23	0.15
Breast Tissue[9x106]	2.47	7.93	2.47	7.93

7. Conclusion

This paper proposes a performance comparison using principal component analysis (PCA) and differential evolution (DE) on fuzzy clustering. PCA, the feature reduction method, is employed as a preprocessing of FCM and KHM. The feature reduction relieves the curse of high-dimensional, noisy data. DE, on the other hand optimizes the initial clusters of FCM and KHM. The aim of such an optimization is to alleviate local optima. This leads to improved FCM and KHM clustering. The comparison on the consequences of feature reduction and those of optimization on the clustering environment is the focus. DEFCM and DEKHM are consecutive manner of outcome is shown in this paper references of DE that optimizes FCM as well as KHM; whereas, FCM and KHM with PCA preprocessing is referred by PCAFCM and PCAKHM. The results, shown in this paper are consistent with those shown in the

related works. (Kao, Lin, & Huang, 2008) denoted better performance of FCM optimized by DE over FCM; the same manner is denoted in this paper.
In addition, this paper as well as (Tian, Liu, & Qi, 2009) point out the better quality yielded by KHM with DE optimization compared to KHM. The application of library evaluation (Wei & Li, 2009) shows the superior results of the algorithm based on PCA and FCM over FCM. Likewise, the same manner of outcome is shown in this paper.

However a comparison study using principal component analysis (PCA) and differential evolution (DE) on fuzzy clustering has not yet been explored in this paper. For this reason, here PCAFCM and PCAKHM are compared with DEFCM and DEKHM. Comparison tests are performed on 7 well-known benchmark real-world data sets. The performance measurements for each method are based on three criteria, FCM / KHM objective functions, ARI and runtime. All of these criteria are calculated with means and standard deviations over 10 independent runs. The results indicate most of superiority of the PCAFCM and PCAKHM over both of DEFCM and DEKHM. The latter optimized algorithms are more complicate and time consuming. Within the scope of this study, one can indicate the superiority of the feature reduction using PCA over DE optimization on FCM and KHM. There still exists some drawbacks to PCAFCM and PCAKHM.

They require a priori known number of clusters, since both of the algorithms are instances of a partitional clustering class. It is not applicable when the number of clusters is unknown. Thereby, other ways of clustering such as agglomerative or divisive should be include in the future.

8. References

Balafar, M. A., Ramli, A. R., Saripan, M. I., Mahmud, R., & Mashohor, S. (2008). Medical image segmentation using fuzzy c-mean (FCM), learning vector quantization (LVQ) and user interaction. *Advanced Intelligent Computing Theories and Applications With Aspects of Contemporary Intelligent Computing Techniques,* 15(5), 177-184.

Bankapalli, J., Venu, B. R., & Devi, A. S. (2011). Combining k-harmonic mean and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging. International Journal on *Computer Science and Engineering, 3*(6), 2544-2553.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences,* 3 (10), 191–203.

Chen, W., Giger, M. L., & Bick, U. (2006). A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Academic Radiology,* 13(1), 63-72.

Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J., & Chen, T.-J. (2006). Fuzzy c-means clustering with spatial information for image segmentation. *Computerized medical imaging and graphics the official journal of the Computerized Medical Imaging Society,* 30(1), 9-15.

Frackiewicz, M. & Palus, H. (2008, December 16-19). Clustering with k-harmonic means applied to colour image quantization. In *Signal Processing and Information Technology, 2008. IEEE International Symposium,* 52-57. Sarajevo, Bosnia and Herzegovena. doi**:**10.1109/ISSPIT.2008.4775684

Frank, A. & Asuncion, A. (2010*). UCI Machine Learning Repository.* School of Information and Computer Science, University of California, Irvine, CA, USA. Retrieved from http://archive.ics.uci.edu/ml

Gomathi, M., & Thangaraj, P. (2010). A parameter based modified fuzzy possibilistic c-means clustering algorithm for lung image segmentation. *Global Journal of Computer Science and Technology, 10*(4), 85-91.

Gungor, Z., & Unler, A. (2007). K-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation,* 184(2), 199-209.

Gungor, Z., & Unler, A. (2008). K-harmonic means data clustering with tabu-search method. *Applied Mathematical Modelling,* 32(6), 1115-1125.

Hamerly, G., & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on*

Information and knowledge management CIKM 02, 600-607. New York, NY, USA.

Jolliffe, I. T. (1986). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1-3), 37-52.

Kannan, S. R., Ramathilagam, S., Sathya, A., & Pandiyarajan, R. (2010). Effective fuzzy c-means based kernel function in segmenting medical images. *Computers in Biology and Medicine,* 40(6), 572-579.

Kao, Y., Lin, J., & Huang, S. (2008). Fuzzy clustering by differential evolution. In *Intelligent Systems Design and Applications, 2008. ISDA '08. Eighth International Conference,* 1*,* 246-250.

Li, D., Gu, H., & Zhang, L. (2010). A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. *Expert Systems with Applications,* 37(10), 6942-6947.

Ma, L., & Staunton, R. (2007). A modified fuzzy c-means image segmentation algorithm for use with uneven illumination patterns. *Pattern Recognition,* 40(11), 3005-3011.

Price, K., Storn, R., & Lampinen, J. (2005). *Differential Evolution: A Practical Approach to Global Optimization*. Book (p. 538). Berlin, Germany: Springer.

Storn, R., & Price, K. (1997). Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization,* 11(4), 341-359.

Tian, Y., Liu, D., & Qi, H. (2009). K-harmonic means data clustering with differential evolution. In *BioMedical Information Engineering, 2009. FBIE 2009. International Conference,* 369-372.

Wang, L., Liu, Y., Zhao, X., & Xu, Y. (2006). Particle swarm optimization for fuzzy c-means clustering. *In Intelligent Control and Automation 2006. WCICA 2006. The Sixth World Congress.* 2*,* 6055-6058.

Wei, Y. & Li, H. (2009). The Library Evaluation Based on the PCA and Fuzzy-c Means. Artificial Intelligence and Computational Intelligence, 2009. *AICI '09. International Conference*, 2, 167-171.

Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization. *Expert Systems with Applications*, 36(6), 9847-9852.

Yong, Y., Chongxun, Z., & Pan, L. (2004). A novel fuzzy c-means clustering algorithm image thresholding. *Measurement Science,* 4(1), 11–19.

Zhang, B. (2000). Generalized k-harmonic means-Boosting in unsupervised learning. *Technical Report TR HPL-2000-137, Hewlett Packard Labs, Palo Alto, CA*.

Zhang, B., Hsu, M., & Dayal, U. (1999). K-harmonic heans -A data clustering algorithm. *Hewllet Packard Research Laboratory Technical Report PL1999124. Citeseer*.