LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction

Cheng-Hsuan Li, Bor-Chen Kuo, *Member, IEEE*, and Chin-Teng Lin, *Fellow, IEEE*

***Abstract*—Research has shown fuzzy c-means (FCM) clustering to be a powerful tool to partition samples into different categories. However, the objective function of FCM is based only on the sum of distances of samples to their cluster centers, which is equal to the trace of the within-cluster scatter matrix. In this study, we propose a clustering algorithm based on both within- and between-cluster scatter matrices, extended from linear discriminant analysis (LDA), and its application to an unsupervised feature extraction (FE). Our proposed methods comprise between- and within-cluster scatter matrices modified from the between- and within-class scatter matrices of LDA. The scatter matrices of LDA are special cases of our proposed unsupervised scatter matrices. The results of experiments on both synthetic and real data show that the proposed clustering algorithm can generate similar or better clustering results than 11 popular clustering algorithms: K-means, K-medoid, FCM, the Gustafson–Kessel, Gath–Geva, possibilistic c-means (PCM), fuzzy PCM, possibilistic FCM, fuzzy compactness and separation, a fuzzy clustering algorithm based on a fuzzy treatment of finite mixtures of multivariate Student's *t* distributions algorithms, and a fuzzy mixture of the Student's *t* factor analyzers model. The results also show that the proposed FE outperforms principal component analysis and independent component analysis.**

***Index Terms*—Cluster scatter matrices, clustering, linear discriminant analysis (LDA), unsupervised feature extraction (FE).**

I. Introduction

CLUSTERING analysis is a tool that assesses the relationships among samples of a dataset by organizing the patterns into different groups, such that patterns within one group show greater similarity to each other than those belonging to different groups [1]. Clustering analysis has the potential to detect underlying structures within data, for classification and pattern recognition, as well as for model reduction and optimization [2], [8].

Manuscript received October 23, 2009; revised March 10, 2010; accepted September 17, 2010. Date of publication October 25, 2010; date of current version February 7, 2011. This work was supported in part by the National Science Council under Grant NSC 99-2221-E-142-002.

C.-H. Li is with the Institute of Electrical Control Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan, and also with the Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung 40306, Taiwan (e-mail: ChengHsuanLi@gmail.com).

B.-C. Kuo is with the Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung 40306, Taiwan (e-mail: kbc@mail.ntcu.edu.tw).

C.-T. Lin is with the Brain Research Center and the Institute of Electrical Control Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: ctlin@mail.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2010.2089631

Fuzzy c-means (FCM) is one of the most well-known clustering methods [3], [4]. However, the objective function of FCM is only based on the sum of distances between samples to their cluster centers, which is equal to the trace of the within-cluster scatter matrix [5], [26]. In recent years, linear discriminant analysis (LDA) [5] has often been used for dimensional reduction in supervised classification problems. LDA uses the mean vector and covariance matrix of each class to formulate within-class, between-, and mixture-class scatter matrices. Based on the Fisher criterion, the LDA method finds features such that the ratio of the between-class scatter to the average within-class scatter is maximized in a lower dimensional space. By applying the concept of class scattering to class separation, the Fisher criterion takes the large values from samples when they are well clustered around their mean within each class, and the clusters of the different classes are well separated [8]. It is formulated as a function of class statistics. For these reasons, we have proposed a clustering algorithm based on the Fisher criterion.

Despite the wide use of FCM-type clustering algorithms, their performance suffers from the curse of dimensionality [34]. It has been shown that using dimension reduction in preprocessing can improve the performance of clustering [28], [45]–[47]. Based on LDA, we have also proposed an unsupervised dimension reduction algorithm in this study.

This paper is organized as follows: Section II introduces reviews of certain popular clustering algorithms and dimension reduction methods. Section III discusses the proposed clustering algorithm and its application to an unsupervised feature extraction (FE) method, and the connection of LDA and the methods to be proposed are shown. In Section IV, both synthetic and real-data experiments are designed to evaluate the performances of the proposed methods, and results of experiments are reported. Section V contains comments and conclusions. Section VI discusses future work.

II. Reviews of Certain Popular Clustering Algorithms and Dimension Reduction Methods

Clustering partitions data into dissimilar groups of similar items and is perhaps the most important and widely used method of unsupervised learning [8]. To avoid the curse of dimensionality (Hughes phenomenon), dimension reduction methods are usually used to discover suitable representations by using specific criteria. In this section, some well-known clustering algorithms and dimension reduction methods are reviewed.

A. Some Popular Clustering Algorithms

1) K-Means and K-Medoid Algorithms: To partition samples into L clusters, two hard clustering methods, i.e., the

K-means (KMS) and K-medoid (KMD) clustering algorithms, are simple and popular [2], [6]–[8]. Unfortunately, their performance is not always reliable or very sensitive to initial clustering centers. KMS and KMD find the centers $c_i, i = 1, \ldots, L$ to minimize the within-cluster sum of squares:

$$J_{\mathrm{KM}} = \sum_{i=1}^{L} \sum_{x_j \in H_i} \|x_j - c_i\|^2$$

where H_i is a set of samples in the ith cluster, and c_i is the center of cluster i.

In KMS, the cluster centers are defined by

$$c_i = \frac{1}{N_i} \sum_{x_j \in H_i} x_j$$

where N_i is the number of samples in H_i. The main difference between KMS and KMD is in calculating the cluster centers: The new cluster center in KMD is the nearest data point to the mean of the cluster points, i.e.,

$$c_i = \arg\min_{x_j \in H_i} \left\| x_j - \frac{1}{N_i} \sum_{x_j \in H_i} x_j \right\|.$$

2) Fuzzy C-Means Clustering Algorithm: FCM clustering is the fuzzy equivalent of the nearest mean "hard" clustering algorithm [2], [6]–[9] that minimize the cost function

$$J_{\mathrm{FCM}}(u_{ij}, c_i) = \sum_{i=1}^{L} \sum_{j=1}^{N} (u_{ij})^m \|x_j - c_i\|^2$$

with respect to membership grade u_{ij}, where c_i is the center of fuzzy cluster i, N is the number of samples, $L > 1$ is the number of clusters, and $m \in (1, \infty)$ is a weighting exponent.

In the FCM algorithm, the memberships, which were inversely related to the relative distance of x_j to the L cluster centers, $\{c_i\}$ were assigned to x_j. The formulation of criterion J_{FCM} could be regarded as the trace of the fuzzy within-cluster scatter matrix [8]

$$S_{\mathrm{FW}} = \sum_{i=1}^{L} \sum_{j=1}^{N} (u_{ij})^m (x_j - c_i)(x_j - c_i)^T$$

.

The earlier form is similar to the within-class scatter matrix of LDA, i.e., this criterion only considers the within-cluster scatter matrix. A consideration of the within-cluster similarity is the only criterion. Based on the suggestion from [52], the division into clusters should be characterized by within-cluster similarity and between-cluster (external) dissimilarity. This is the reason that we applied the Fisher criterion in this study.

3) Fuzzy Compactness and Separation: Two similar fuzzy-based clustering algorithms based on fuzzy within-cluster, between-cluster, and total scatter matrices are proposed in [26] and [42]. The objective function of the fuzzy compactness and separation (FCS) [26] is based on fuzzy between- and within-cluster scatter matrices. Hence, the measurement of compactness is minimized, and the separation measure can be maximized simultaneously.

The fuzzy between-cluster scatter matrix S_{FB} and within-cluster scatter matrix S_{FW} are defined as follows:

$$S_{\mathrm{FB}} = \sum_{i=1}^{c} \sum_{j=1}^{n} \eta_i (u_{ij})^m (x_j - c)(x_j - c)^T$$

and

$$S_{\mathrm{FW}} = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (x_j - c_i)(x_j - c_i)^T$$

where $c = 1/N \sum_{i=1}^{N} x_j$. The objective function of FCS is defined as follows:

$$\begin{aligned} J_{\mathrm{FCS}}(u_{ij}, c_i) &= \mathrm{tr}(S_{\mathrm{FW}}) - \mathrm{tr}(S_{\mathrm{FB}}) \\ &= \sum_{i=1}^{L} \sum_{j=1}^{N} (u_{ij})^m \|x_j - c_i\|^2 \\ &\quad - \sum_{i=1}^{L} \sum_{j=1}^{N} \eta_i (u_{ij})^m \|x_j - c\|^2. \end{aligned}$$

By minimizing J_{FCS}, we use the following equations that mutually update each other:

$$u_{ij} = \frac{(\|x_j - c_i\|^2 - \eta_i \|c_i - c\|^2)^{-1/(m-1)}}{\sum_{k=1}^{L} (\|x_j - c_k\|^2 - \eta_k \|c_k - c\|^2)^{-1/(m-1)}}$$

and

$$c_i = \frac{\sum_{j=1}^{N} (u_{ij})^m x_j - \eta_i \sum_{j=1}^{n} (u_{ij})^m c}{\sum_{j=1}^{N} (u_{ij})^m - \eta_i \sum_{j=1}^{n} (u_{ij})^m}$$

where the parameter η_i could be set up with $\eta_i = ((\beta/4) \min_{i' \neq i} \|c_i - c_{i'}\|^2) / \max_k \|c_k - c\|^2$, and $\beta \in [0, 1]$ is the parameter that should be predetermined. Note that the objective function proposed by Yin *et al.* [42] is a special case of FCS when the parameters η_i are all set to $1/(L(L-1))$.

The Fisher criterion, $\mathrm{tr}[(S_{\mathrm{FW}})^{-1} S_{\mathrm{FB}}]$ takes large values when samples are well clustered around their mean within each class, and the clusters of the different classes are well separated [8]. This approach is widely used in different applications [33], [43], [44]. In the following, the new definitions of unsupervised cluster scatter matrices are introduced, and the corresponding objective function is based on the Fisher criterion, including the interaction of cluster scatter matrices.

4) Other Advanced FCM-Type Clustering Algorithms: A typical well-known algorithm of this category is the Gustafson–Kessel (GK) algorithm [29], which employs an adaptive distance norm, to detect clusters of different geometrical shapes in one dataset [2]. In addition, Krishnapuram and Keller [24] proposed a new clustering model named possibilistic c-means (PCM), which relaxes the constraint "the sum of the membership values of every sample to all clusters is" in order to interpret in a possibilistic sense the membership function or degree of typicality [53]. In 1997, the fuzzy PCM (FPCM) [31] was proposed to generate both possibility and membership values. However, the possibility values generated by FPCM are very small as the size of the dataset increases. To eliminate the problem of FPCM and integrate the benefits of FCM and PCM, the possibilistic FCM (PFCM) was proposed in 2005 [25].

Some FCM-type algorithms employ an adaptive distance norm based on the fuzzy maximum likelihood estimates [2], [30], such as the Gath–Geva (GG) algorithm. Chatzis and Varvarigou [27] proposed a robust fuzzy clustering algorithm based on a fuzzy treatment of finite mixtures of multivariate Student's *t* distributions (FSMM), which used finite mixtures of multivariate Student's *t* distributions, instead of finite Gaussian mixture models (GMMs).

B. Reviews of Dimension Reduction Methods

There are some main divisions for dimension reduction, e.g., filters [35], [36], wrappers [37], [38], embedded maps [39], [40], and hybrid search [41]. In the classical statistical method, principal component analysis (PCA) [12], [13] and independent component analysis (ICA) [14] are commonly applied to dimension reduction. Factor analysis is a linear latent variable scheme, which is used to capture local substructures and local dimensionality reduction techniques. Chatzis and Varvarigou [28] combined the advantages of factor analysis and proposed a fuzzy mixture of Student's *t* factor analyzers (FMSFA). FMSFA provided a well-established observation space dimensionality reduction framework for fuzzy clustering algorithms based on factor analysis. This allows concurrent performance of fuzzy clustering and, within each cluster, a reduction in local dimensionality. Their experimental results show that FMSFA outperforms finite mixtures of Student's *t* factor analyzers (*t*MFA) [46], the KLFCV model of Honda and Ichihashi [49], and the MFA model [50]. In this study, the performance of the proposed FE will be compared with those of PCA, ICA, and FMSFA.

LDA is often used for dimension reduction in classification problems. It is often referred to the parametric FE method in [5], since LDA uses the mean vector and covariance matrix of each class. Usually, within-class, between class, and mixture scatter matrices are used to formulate the criterion of class separability.

Suppose that $H_i = \{x_1^{(i)}, \ldots, x_{N_i}^{(i)}\} \subset R^d$ were the set of samples in class i, N_i were the number of samples in class i, $i = 1, \ldots, L$, and $N = N_1 + \cdots + N_L$ were the number of all training samples. In LDA, the between-class scatter matrix S_b^{LDA} and the within-class scatter matrix S_w^{LDA} would be defined as

$$S_b^{\mathrm{LDA}} = \sum_{i=1}^{L} N_i/N (c_i - c)(c_i - c)^T$$

and

$$S_w^{\mathrm{LDA}} = \sum_{i=1}^{L}\sum_{j=1}^{N_i} \frac{1}{N}\left(x_j^{(i)} - c_i\right)\left(x_j^{(i)} - c_i\right)^T$$

where c_i were the class mean defined by $c_i = 1/N_i \sum_{j=1}^{N_i} x_j^{(i)}$ and $c = 1/N \sum_{i=1}^{L}\sum_{j=1}^{N_i} x_j^{(i)}$ representing the total mean.

The optimal features are determined by optimizing the Fisher criterion $J_{\mathrm{LDA}} = J_1$ given by

$$J_{\mathrm{LDA}} = \mathrm{tr}\left[\left(S_w^{\mathrm{LDA}}\right)^{-1} S_b^{\mathrm{LDA}}\right].$$

This is equivalent to solving the generalized eignevalue problem

$$S_b^{\mathrm{LDA}} v_s = \lambda_s S_w^{\mathrm{LDA}} v_s, s = 1, \ldots, d \text{ with } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

where the extracted eigenvectors are used to form the transformation matrix of LDA, i.e., the transformation matrix from the original space to the reduced subspace is defined by

$$A = [v_1, v_2, \ldots, v_p].$$

The Fisher criterion J_{LDA} is able to detect the capacity of the separability for the transformed training samples but LDA is a supervised FE. In Section III, we propose the between- and within-cluster scatter matrices of an unsupervised LDA (UFLDA) using the concept of membership values and cluster means of FCM as a clustering algorithm and an unsupervised FE.

III. Proposed Method: Linear-Discriminant-Analysis-Based Clustering Algorithm and Unsupervised Linear Discriminant Analysis

In this section, we propose a novel clustering algorithm, i.e., LDA-based clustering (FLDC), and its application to an unsupervised FE and UFLDA.

Two unsupervised between- and within-cluster scatter matrices are derived from the scatter matrices of LDA and are applied to formulate FLDC. We define the between-cluster scatter matrix S_b^{UFLDA} and the within-cluster scatter matrix S_w^{UFLDA} as follows:

$$S_b^{\mathrm{UFLDA}} = \sum_{i=1}^{L} \frac{\sum_{j=1}^{N} u_{ij}}{N}(c_i - c)(c_i - c)^T$$

and

$$S_w^{\mathrm{UFLDA}} = \sum_{i=1}^{c}\sum_{j=1}^{N} \frac{u_{ij}}{N}(x_j - c_i)(x_j - c_i)^T$$

where $c_i = \sum_{j=1}^{N} (u_{ij}/\sum_{k=1}^{N} u_{ik}) x_j$ is the class mean, which is the same as FCM, and $c = 1/N \sum_{i=1}^{N} x_j$ represents the total mean. The following theorem shows that the between- and within-class scatter matrices of LDA are special cases of our proposed S_b^{UFLDA} and S_w^{UFLDA}, respectively.

Theorem 1: In the supervised situation, if

$$u_{ij} = \begin{cases} 1, & \text{if } x_j \in H_i \\ 0, & \text{if } x_j \notin H_i \end{cases} \quad \text{for all } 1 \leq i \leq L \text{ and } 1 \leq j \leq N$$

then the proposed S_b^{UFLDA} and S_w^{UFLDA} are the same as S_b^{LDA} and S_w^{LDA}, respectively.

Proof: Suppose there are N_i samples in H_i for $i = 1, \ldots, L$ and $\sum_{k=1}^{N} u_{ik} = N_i$. Then

$$c_i = \sum_{j=1}^{N} \frac{u_{ij}}{\sum_{k=1}^{N} u_{ik}} x_j = \sum_{x_j \in H_i} \frac{1}{N_i} x_j = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j^{(i)}$$

is the same as the class mean shown in LDA and the between-cluster scatter matrix

$$S_b^{\mathrm{UFLDA}} = \sum_{i=1}^{L} \frac{\sum_{j=1}^{N} u_{ij}}{N}(c_i - c)(c_i - c)^T$$
$$= \sum_{i=1}^{L} \frac{N_i}{N}(c_i - c)(c_i - c)^T = S_b^{\mathrm{LDA}}.$$

The within-cluster scatter matrix is, thus, as follows:

$$S_w^{\mathrm{UFLDA}} = \sum_{i=1}^{c}\sum_{j=1}^{N} \frac{u_{ij}}{N}(x_j - c_i)(x_j - c_i)^T$$
$$= \sum_{i=1}^{c}\sum_{x_j \in H_i} \frac{u_{ij}}{N}(x_j - c_i)(x_j - c_i)^T$$
$$= \sum_{i=1}^{c}\sum_{j=1}^{N_i} \frac{1}{N}\left(x_j^{(i)} - c_i\right)\left(x_j^{(i)} - c_i\right)^T = S_w^{\mathrm{LDA}}.$$

Based on this theorem and the objective function of LDA, the general objective function of FLDC is defined by

$$J_{\mathrm{FLDC}}(u_{ij}) = \mathrm{tr}\left[\left(S_w^{\mathrm{UFLDA}}\right)^{-1} S_b^{\mathrm{UFLDA}}\right]$$

including the interaction of S_b^{UFLDA} and S_w^{UFLDA}. We consider the interaction of the between- and within-cluster scatter matrices in the Fisher criterion. In our results on the artificial datasets, FLDC is able to detect the clusters with the largest between-cluster separability.

To reduce the effect of the cross products of within-class distances and prevent singularity, some regularized techniques [15], [16] can be applied to within-cluster scatter matrix. In FLDC, the within-cluster scatter matrix is regularized by

$$S_{rw}^{\mathrm{UFLDA}} = rS_w^{\mathrm{UFLDA}} + (1 - r)\mathrm{diag}\left(S_w^{\mathrm{UFLDA}}\right)$$

where $\mathrm{diag}(S_w^{\mathrm{UFLDA}})$ is the diagonal parts of matrix S_w^{UFLDA}, and $r \in [0, 1]$ is a regularization parameter.

In the proposed clustering algorithm, the optimization problem is defined as follows:

$$U_{\mathrm{FLDC}} = \arg\max_{U} J_{\mathrm{FLDC}}(u_{ij})$$
$$= \arg\max_{U} \left[\left(S_{rw}^{\mathrm{UFLDA}}\right)^{-1} S_b^{\mathrm{ULDA}}\right]$$

which constrains $\sum_{i=1}^{L} u_{ij} = 1, j = 1, \ldots, N$. Since the optimization problem is nonlinear and nonconvex, some popular optimization algorithms [17], [20]: "interior-point," "active-set," and "trust-region-reflective" can be applied to solve this problem. In implementing these algorithms, we find that the cost in time of the "active-set" algorithm is less than the other two algorithms, but it is sensitive to the initial value. Hence, the "interior-point" algorithm is used to find the optimizer U_{FLDC}. However, the corresponding time cost of the "interior-point" algorithm is higher than the others.

After finding the optimizer U_{FLDC} in FLDC, the features of UFLDA can be obtained by solving the following generalized eignevalue problem:

$$S_b^{\mathrm{UFLDA}} v_s = \lambda_s S_{rw}^{\mathrm{UFLDA}} v_s, s = 1, \ldots, d$$

with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

where the elements u_{ij} is in U_{FLDC}. The extracted eigenvectors form the transformation matrix of UFLDA. The algorithm of UFLDA is described as follows:

Step 1: Initialize the membership matrix $U = [u_{ij}]_{1 \leq i \leq L, 1 \leq j \leq N}$ with random values from $[0, 1]$ such that the element u_{ij} of U satisfies $\sum_{i=1}^{L} u_{ij} = 1, j = 1, \ldots, N$.

Step 2: Use the interior-point optimization method to find the optimal U_{FLDC} under the Fisher criterion $\max_U \mathrm{tr}[(S_{rw}^{\mathrm{UFLDA}})^{-1} S_b^{\mathrm{UFLDA}}]$, which constrains

$$\sum_{i=1}^{L} u_{ij} = 1, j = 1, \ldots, N$$

and

$$u_{ij} \in [0, 1], i = 1, \ldots, L, j = 1, \ldots, N.$$

Step 3: Compute S_{rw}^{UFLDA} and S_b^{UFLDA} with the optimal U_{FLDC} obtained in Step 2.

Step 4: Solve the eignevalue problem $(S_{rw}^{\mathrm{UFLDA}})^{-1} S_b^{\mathrm{UFLDA}} v_s = \lambda_s v_s$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

Step 5: The transformation function from the original space to the reduced subspace is defined by $A = [v_1, v_2, \ldots, v_p], p \leq d$.

IV. Experiments

The performances of the proposed FLDC and UFLDA were validated in Experiments 1 and 2, respectively. We selected ten artificial and three real datasets for the experiments. We compared the results of clustering FLDC, KMS, KMD, FCM, GK algorithm, GG algorithm [2], PCM [24], FPCM [31], PFCM [25], FCS [26], FSMM [27], and FMSFA [28], [51] algorithms on artificial and real datasets. The parameters r in FLDC and β in FCS were set to 0.5. The weighting exponents of FCM, GK, GG, and PCM were set to $m \in \{2, 4\}$. The weighting exponents of FPCM and PFCM were set to $m \in \{2, 4\}$ and $\eta \in \{2, 4\}$. The parameters of FSMM were set to the default values [51]. The results of clustering FMSFA were the best results within the given set $\{0.5, 1, 1.5\}$ of the model's degrees of fuzziness of the fuzzy membership values.

In order to avoid the influence of initialization, all clustering algorithms were evaluated, based on the three real datasets and 100 randomly generated initial values for each dataset. The mean, standard deviation, maximum, and minimum accuracy of the 100 clustering accuracy are calculated and compared. The accuracy of the clustering is the proportion of correctly clustered data in the dataset, i.e., clustering accuracy = (the number of correctly clustered data)/(the number of all samples).

In Experiment 2, the scatter plots of transformed data with two-dimensional (2-D) extracted features obtained from UFLDA, PCA, and ICA are displayed. We compare the accuracy of the clustering obtained from the first ten clustering algorithms with extracted features in this experiment.

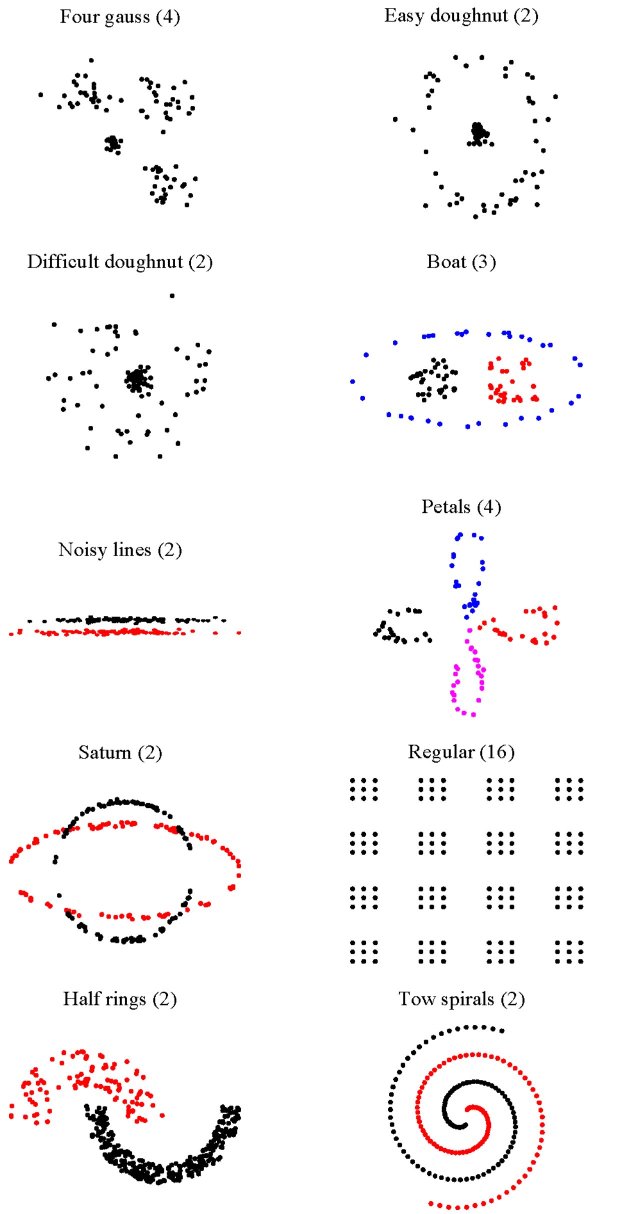


Fig. 1. Ten artificial datasets [22] were used in this study. The first three datasets were generated with ten additional noise features. The number of clusters is given in parentheses.

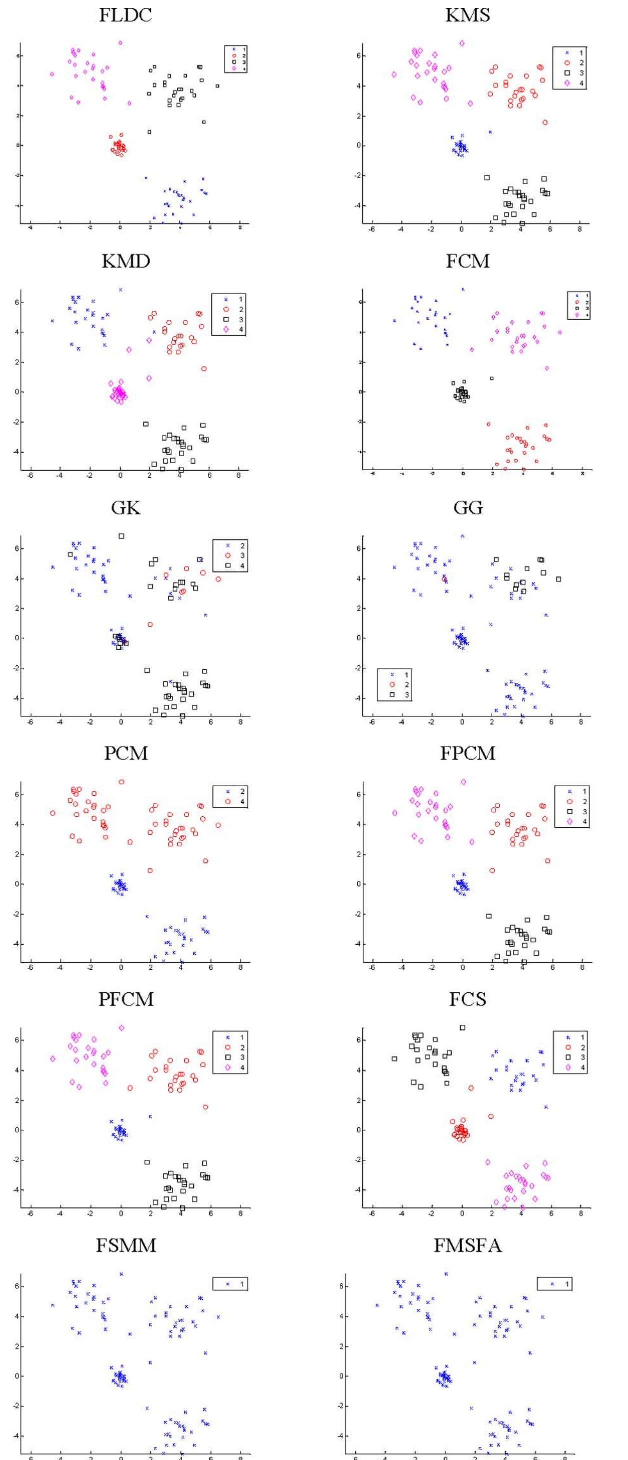


Fig. 2. Results of clustering the "Four gauss" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

TABLE I
DESCRIPTIONS OF THREE REAL DATASETS

Dataset	Classes	# of Samples	Features
Wine	3	178	13
Iris	3	150	4
WDBC	2	569	30

A. Datasets

Ten artificial datasets [21] containing "Four-gauss data" (four clusters), "Easy doughnut data" (two clusters), "Difficult doughnut data" (two clusters), "Boat data" (three clusters), "Noisy lines data" (two clusters), "Petals data", (four clusters), "Saturn data" (two clusters), "Regular data" (16 clusters), "Half-ring data" (two clusters), and "Spirals data" (two clusters) and can be downloaded from [22] and are presented in Fig. 1. These were all created in two dimensions and are meant to present challenges in varying degrees to the clustering algorithm. Ten dimensions of uniformly random noise were appended to each of the first three datasets (four gauss, easy doughnut, and difficult doughnut), while the other seven datasets were kept as 2-D.

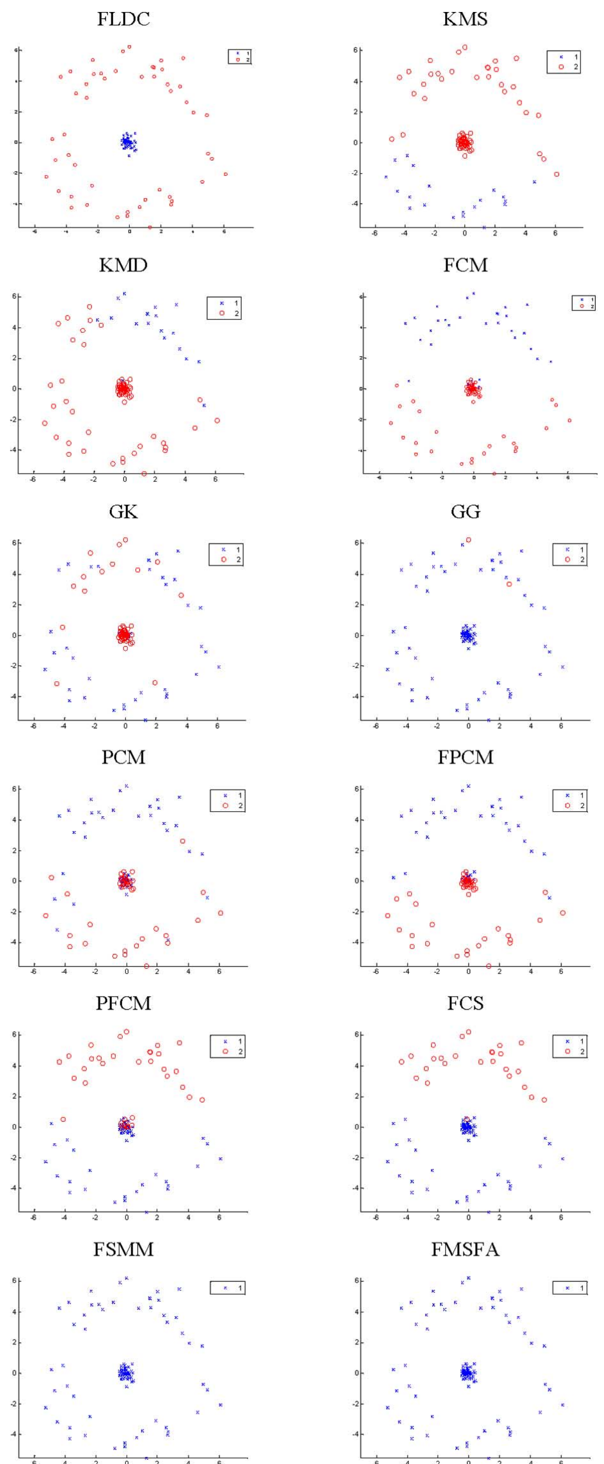


Fig. 3. Results of clustering "Easy doughnut" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

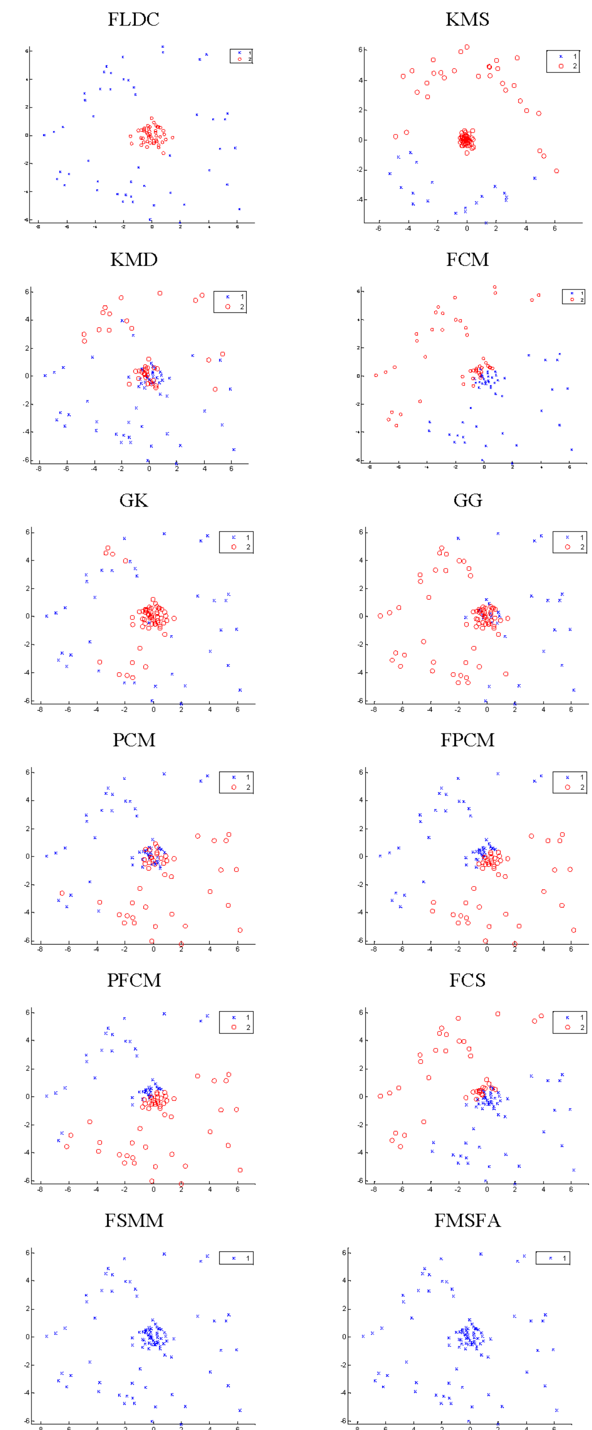


Fig. 4. Results of clustering "Difficult doughnut" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

The last two datasets were omitted because the clustering results obtained by applying all clustering algorithms are similar.

The real datasets, "Wine," "Iris," and "Breast Cancer Wisconsin (Diagnostic)" (WDBC) are described in Table I. The wine dataset is a collection of data using three classes of wine from various locations in Italy. Iris dataset contains three classes, Iris Setosa, Iris Versicolour, and Iris Virginica of Iris flowers collected from Hawaii. There are two classes, i.e., benign and malignant, in the WDBC dataset. These datasets are available from the FTP server of the University of California at Irvine (UCI) [23] data repository.

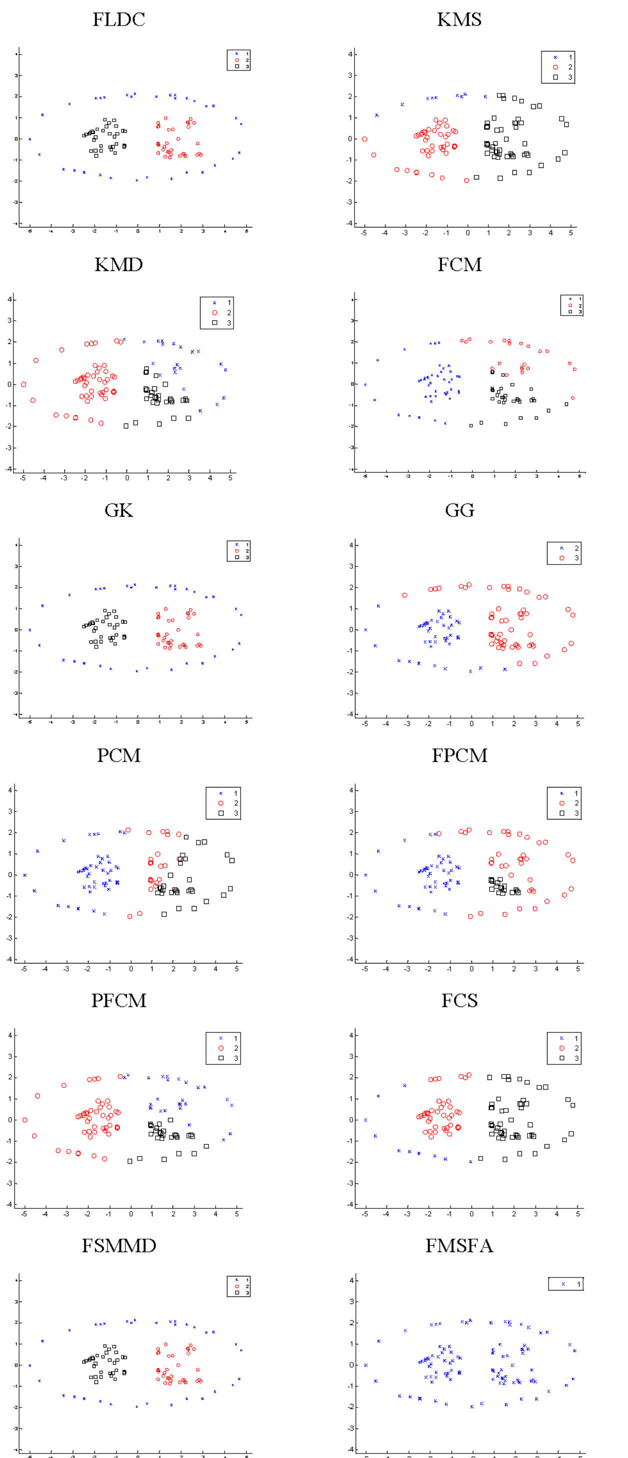


Fig. 5. Results of clustering "Boat" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

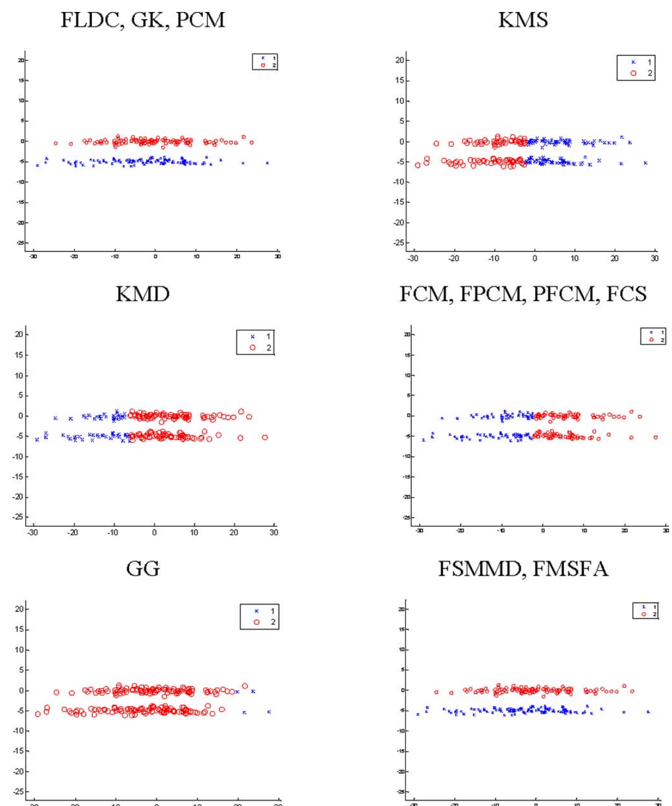


Fig. 6. Results of clustering "Noisy lines" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

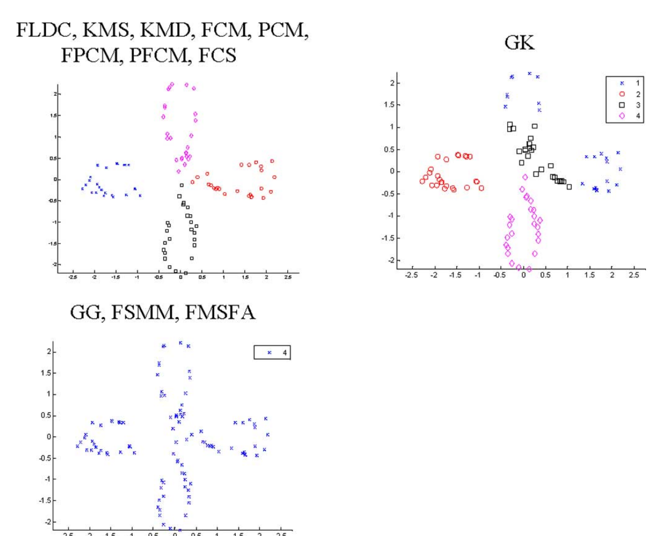


Fig. 7. These figures show the results of clustering "Petals" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

B. *Results of Experiment 1*

The results of clustering Experiment 1 on the artificial datasets are shown in Figs. 2–9. The covariance matrices of two density-based methods, i.e., GG and FSMM, are near-singular. Hence, we used the GG and FSMM with diagonal covariance matrices for the Gaussian distributions (GGD) and the Student's *t* distributions (FSMMD), respectively. The best clustering results from the application of GG and GGD in different datasets were chosen for Figs. 2–9. The best results of clustering FSMM and FSMMD were also shown in Figs. 2–9. In comparing Figs. 2–9, we find the following:

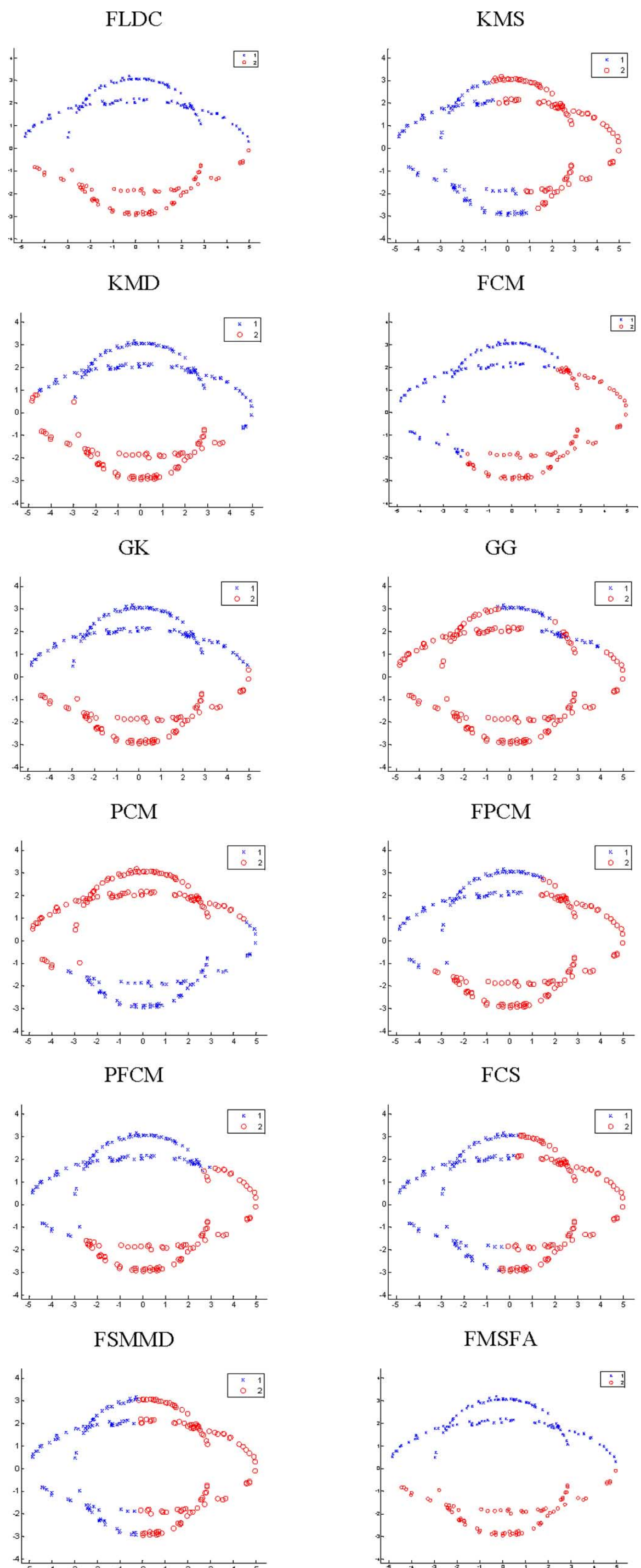


Fig. 8. Results of clustering "Saturn" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

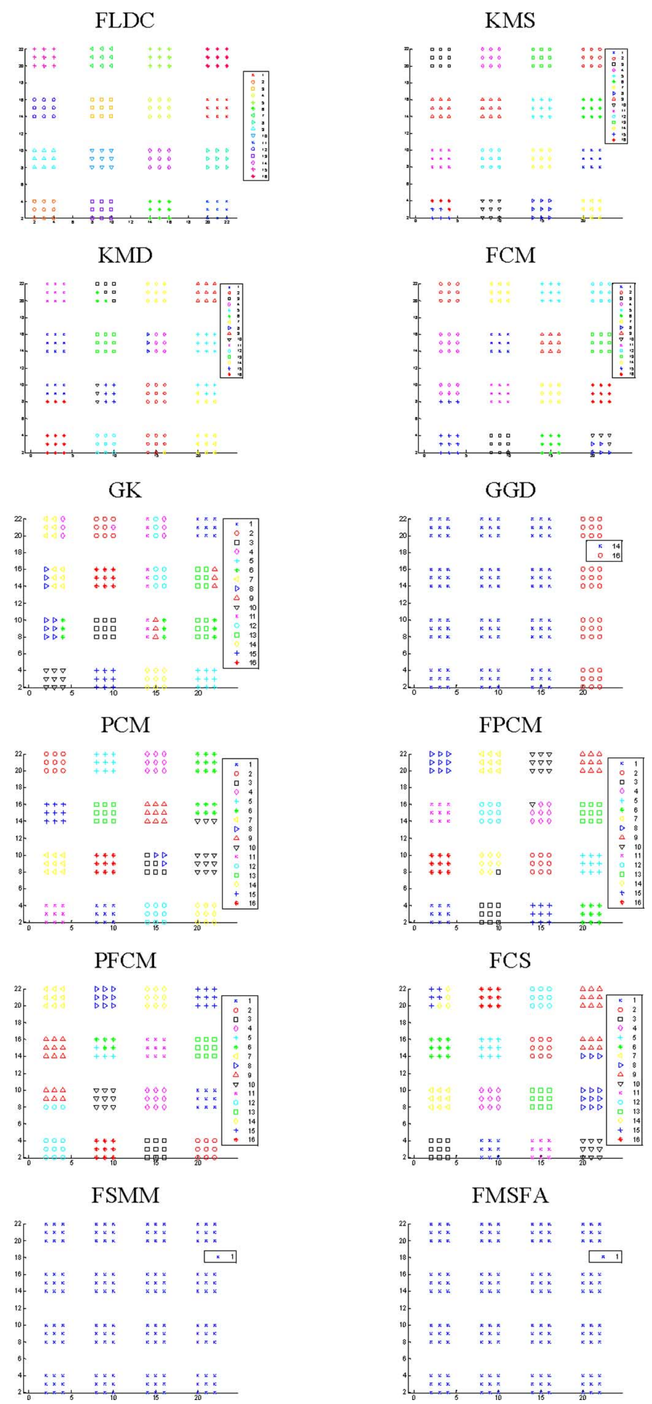


Fig. 9. Results of clustering "Regular" dataset applied by 12 clustering algorithms. The best clustering results from the application of GG and GGD were chosen for comparison. Similarly, the best results of clustering FSMM and FSMMD were also shown for comparison.

1) The results of clustering FLDC significantly outperformed or were equal to others for the normal-like distribution of data, for instance, four-gauss, easy doughnut, difficult doughnut, boat, and petals datasets, because the interaction of the between- and within-cluster scatter matrices was considered when applying FLDC. Especially in the easy doughnut and difficult doughnut, all algorithms had poor clustering results, except for FLDC.

TABLE II
MEAN, STANDARD DEVIATION, MAXIMUM, AND MINIMUM ACCURACY OF CLUSTERING OF THREE REAL DATASETS

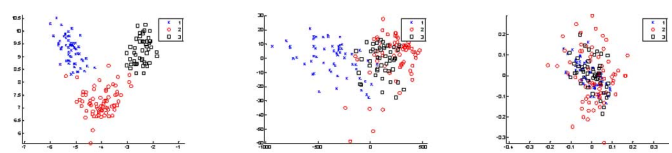
Method \ Dataset	Wine mean/std/max/min (mean of cpu time in sec.)	Iris mean/std/max/min (mean of cpu time in sec.)	WDBC mean/std/max/min (mean of cpu time in sec.)
FLDC	**0.972**/0.003/0.949/0.916 (106.628)	**0.966**/0.001/0.967/0.960 (43.871)	**0.940**/0.001/0.946/0.938 (2099.937)
KMS	0.677/0.051/0.702/0.567 (0.006)	0.849/0.109/0.893/0.580 (0.009)	0.854/0.000/0.854/0.854 (0.009)
KMD	0.667/0.062/0.708/0.556 (0.004)	0.835/0.141/0.947/0.513 (0.006)	0.851/0.006/0.854/0.837 (0.008)
FCM	0.691/0.000/0.691/0.691 (0.119)	0.907/0.000/0.907/0.907 (0.002)	0.861/0.000/0.861/0.861 (0.108)
GK	0.607/0.000/0.607/0.607 (1.726)	0.900/0.000/0.900/0.900 (0.058)	0.821/0.000/0.821/0.821 (0.254)
GG	0.742/0.000/0.742/0.742 (0.218)	0.733/0.000/0.733/0.733 (0.135)	0.510/0.000/0.510/0.510 (0.113)
PCM	0.697/0.000/0.697/0.697 (0.015)	0.933/0.000/0.933/0.933 (0.014)	0.856/0.000/0.856/0.856 (0.033)
FPCM	0.719/0.000/0.719/0.719 (0.027)	0.907/0.000/0.907/0.907 (0.017)	0.877/0.000/0.877/0.877 (0.036)
PFCM	0.691/0.000/0.691/0.691 (0.060)	0.920/0.000/0.920/0.920 (0.027)	0.861/0.000/0.861/0.861 (0.046)
FCS	0.697/0.000/0.697/0.697 (0.249)	0.893/0.000/0.893/0.893 (0.112)	0.851/0.000/0.851/0.851 (0.321)
FSMM	0.846/0.117/0.899/0.573 (0.583)	0.875/0.170/0.973/0.527 (0.066)	0.935/0.000/0.935/0.935 (0.299)

2) The FLDC had the best performance with the regular and noisy lines datasets.
3) KMS, KMD, FCM, FPCM, PFCM, and FCS performed well only on the four gauss and petals datasets.
4) PCM performs well only on the petals and noisy lines datasets.
5) GK employed an adaptive norm by estimating covariance matrices for each cluster. Hence, it could detect clusters with different geometrical shapes, and it performed well on the boat and noisy lines datasets. However, on the four gauss, easy doughnut, and difficult doughnut datasets, the performance was dismal.
6) Although FLDC performed poorly on the saturn, half rings, and tow spirals datasets, it was able to detect the clusters with the largest between-cluster separability in the saturn dataset. FLDC was unsuitable for saturn, half rings, and tow spirals data as these were complex nonlinear problems. Perhaps, the kernel method would be a way to solve these types of datasets.
7) The distribution-based clustering algorithms GG, FSMM, and FMSFA performed poorly on the four gauss, easy doughnut, petals, and regular datasets, because the covariance matrices of the density-based methods are near-singular.
8) FSMMD was able to perform better by applying FSMM on the boat and noisy lines.

Table II shows the clustering accuracy in Experiment 1 using real datasets. Note that the highest mean clustering accuracy for each dataset (in rows) is shaded. From Table II, we can see that the highest mean accuracy among all methods was 0.972, 0.966, and 0.940. All of these results were obtained by performing FLDC.

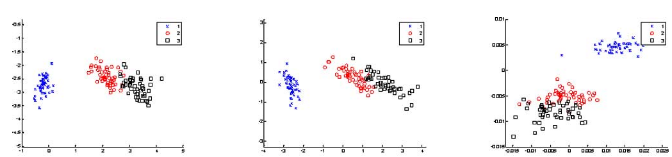
C. Results of Experiment 2

The scatter plots of transformed data with 2-D extracted features obtained from UFLDA, PCA, and ICA on three real datasets are displayed in Fig. 10–12. Moreover, the accuracy of the three real datasets after applying FMSFA is shown in the Table III. The maximum accuracy of these datasets was 1, 0.980, and 0.949, respectively. However, it is very sensitive to the initial value. Hence, the highest average accuracy in every column was only 0.945, 0.774, and 0.882. Because the FSMM



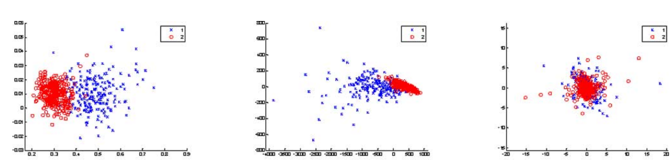
(a) (b) (c)

Fig. 10. Two-dimensional projections of “Wine” data by (a) UFLDA, (b) PCA, and (c) ICA, respectively. The legends 1, 2, and 3 indicate three classes of wine from various locations in Italy.



(a) (b) (c)

Fig. 11. Two-dimensional projections of “Iris” data by (a) UFLDA, (b) PCA, and (c) ICA, respectively. The legends 1, 2, and 3 indicate “Iris Setosa,” “Iris Versicolour,” and “Iris Virginica,” respectively.



(a) (b) (c)

Fig. 12. Two-dimensional projections of “WDBC” data by (a) UFLDA, (b) PCA, and (c) ICA, respectively. The legends 1 and 2 indicate “malignant” and “benign,” respectively.

TABLE III
MEAN, STANDARD DEVIATION, MAXIMUM, AND MINIMUM ACCURACY OF CLUSTERING OF THREE REAL DATASETS BY APPLYING FMSFA, WHERE LD REPRESENTS THE LATENT DIMENSION

Method \ Dataset	Wine	Iris	WDBC
	mean/std/max/min	mean/std/max/min	mean/std/max/min
FMSFA LD=1	0.898/0.085/0.955/0.854	0.774/0.154/0.980/0.333	0.813/0.005/0.821/0.803
FMSFA LD=2	0.945/0.069/0.966/0.579	0.768/0.127/0.967/0.333	0.882/0.000/0.882/0.882
FMSFA LD=3	0.891/0.143/1.000/0.539	0.704/0.105/0.967/0.333	0.865/0.027/0.949/0.715

TABLE IV
HIGHEST MEAN ACCURACY AND THE NUMBERS OF USED FEATURES (IN BRACKETS) (WINE DATA)

Clustering Method \ FE Method	UFLDA	PCA	ICA
KMS	0.972 (2)	0.702 (3)	0.702 (2)
KMD	**0.978 (2)**	0.708 (2)	0.701 (2)
FCM	0.972 (2)	0.691 (1)	0.691 (1)
GK	0.966 (2)	0.967 (2)	0.697 (2)
GG	0.910 (1)	0.847 (4)	0.478 (2)
PCM	0.961 (2)	0.697 (2)	0.719 (2)
FPCM	0.961 (2)	0.719 (1)	0.719 (2)
PFCM	0.966 (2)	0.691 (1)	0.685 (1)
FCS	0.972 (2)	0.697 (1)	0.697 (1)
FSMM	0.972 (2)	0.780 (4)	0.725 (1)

TABLE V
HIGHEST MEAN ACCURACY AND THE NUMBER OF USED FEATURES (IN BRACKETS) (IRIS DATA)

Clustering Method \ FE Method	UFLDA	PCA	ICA
KMS	0.960 (2)	0.913 (1)	0.893 (3)
KMD	0.973 (1)	0.913 (1)	0.896 (3)
FCM	0.974 (1)	0.913 (1)	0.913 (3)
GK	0.974 (1)	0.967 (2)	0.973 (2)
GG	0.973 (1)	0.713 (3)	0.693 (3)
PCM	0.940 (1)	0.940 (3)	0.907 (3)
FPCM	0.973 (1)	0.927 (1)	0.933 (3)
PFCM	**0.980 (2)**	0.940 (1)	0.927 (3)
FCS	0.974 (1)	0.913 (1)	0.900 (3)
FSMM	0.973 (1)	0.973 (2)	0.967 (3)

used the results of clustering KMS as the initial value, it was more stable than FMSFA.

After comparing the accuracy of clustering, we compared the capacity for separation of extracted features by UFLDA, PCA, and ICA. The 2-D scatter plots for the "Wine" dataset extracted by UFLDA, PCA, and ICA are displayed in Fig. 10. The purpose of PCA is to reduce dimensionality according to the selected percentage of the overall variance that can be captured, and the purpose of ICA is to find the underlying factors or sources. In the 2-D scatter plot figures, the capacity for separation was poor. However, the proposed UFLDA was based on the Fisher criterion to discover the preferred features for discrimination. We can clearly observe that there were three clusters in the reduced 2-D subspace. We expect that the clustering algorithms performed better on the UFLDA subspace than on the original space, the PCA, or the ICA subspaces.

The highest mean clustering accuracy from extracted features by UFLDA, PCA, and ICA, which were applied to this dataset using ten popular clustering algorithms, is shown in Table IV. We observe that the KMD with UFLDA features obtained the best performance in the wine dataset. From Tables II and IV, the clustering algorithms performed better with UFLDA features than with the original features.

The 2-D scatter plots for the "Iris" dataset extracted by UFLDA, PCA, and ICA are displayed in Fig. 11. Although two classes "versicolor" and "virginica" are overlapped in the dataset, it is observed that the capacity for separation of UFLDA was improved.

Table V shows the highest mean accuracy using extracted features by UFLDA, PCA, and ICA, which were applied to the Iris dataset in ten popular clustering algorithms. It was observed that using extracted features by UFLDA could obtain an improvement in performance. The greatest accuracy was 0.980 (UFLDA + PFCM). In Tables II and V, performing clustering algorithms with UFLDA features may improve the accuracy over that of the original features.

TABLE VI
HIGHEST MEAN ACCURACY AND THE NUMBERS OF USED FEATURES (IN BRACKETS) (WDBC DATA)

Clustering Method \ FE Method	UFLDA	PCA	ICA
KMS	0.940 (1)	0.854 (1)	0.863 (2)
KMD	0.940 (1)	0.854 (1)	0.854 (1)
FCM	0.946 (1)	0.861 (1)	0.852 (1)
GK	0.958 (2)	0.854 (1)	0.852 (1)
GG	**0.960 (1)**	0.935 (2)	0.931 (2)
PCM	0.951 (1)	0.856 (3)	0.859 (1)
FPCM	0.944 (2)	0.900 (2)	0.886 (2)
PFCM	0.951 (1)	0.861 (1)	0.865 (2)
FCS	0.939 (1)	0.851 (2)	0.849 (1)
FSMM	0.944 (1)	0.931 (2)	0.937 (3)

Finally, the 2-D scatter plots for the "WDBC" dataset extracted by UFLDA, PCA, and ICA are displayed in Fig. 12. In spite of the overlapping samples in these two figures, the labeled samples show that the clustering results of UFLDA were superior to those of PCA and ICA. In Table VI, the results are similar to the aforementioned two datasets. It was observed that using the UFLDA features could achieve equal or improved performance in the WDBC dataset. The highest accuracy was 0.960 (UFLDA + GG).

The selected features by UFLDA are of the same order of magnitude as the eigenvalues, i.e., the magnitude of separability of the projected samples. Therefore, the number of features selected by UFLDA was quite small, i.e., only one or two selected features, for all three real datasets.

V. Conclusion

In this study, we proposed a clustering algorithm FLDC based on the Fisher criterion composed of the between- and within-cluster scatter matrices extended from LDA and its application for an unsupervised FE, UFLDA. The results of the experiment with both synthetic and real data have shown that our proposed clustering algorithm outperformed the KMS, KMD, FCM, GK, GG, PCM, FPCM, PFCM, FCS, FSMM, and FMSFA algorithms. We have also observed an improvement in separabilities with the 2-D projections of three real datasets by UFLDA. When we applied UFLDA features to clustering algorithms, the accuracy was higher than that of PCA and ICA features to clustering algorithms.

VI. Future Work

From the results of clustering synthetic datasets, we observe that FLDC only worked well when the distribution of clusters showed normal-like distribution. Hence, in upcoming research, we can extend FLDC by kernel tricks, i.e., a clustering algorithm based on an unsupervised version of kernel-based LDA for nonnormal datasets.

Another upcoming direction in research shows the proposed optimization problem is nonconvex and nonlinear. Although the proposed methods work well, the optimal solution may be a local minimum, and the interior-point optimization method is time consuming. It will be necessary to find a more efficient algorithm to solve such problems.

Finally, the number of clusters is an important aspect for all clustering algorithms. Our next study will be to develop or choose an appropriate criterion for FLDC (Akaike and Bayesian information criteria) to determine the number of clusters.

Acknowledgment

The authors would like to thank Prof. S. Chatzis for the FSMM and FMSFA algorithms (http://web.mac.com/soteri0s/Sotirios_Chatzis/Software.html) and Prof. L. I. Kuncheva for the artificial datasets (http://www.bangor.ac.uk/∼mas00a/). Finally, we also appreciate N. Pal and the reviewers for providing many useful suggestions.

References

[1] L. X. Xuanli and B. Gerardo, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 1991.

[2] B. Balasko, J. Abonyi, and B. Feil, *Fuzzy clustering and data analysis toolbox for use with MATLAB.* (2010). [Online]. Available from: http://www.fmt.vein.hu/softcomp

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[4] P. J. Rouseeuw, L. Kaufman, and E. Trauwaert, "Fuzzy clustering using scatter matrices," *Comput. Statist. Data Anal.*, vol. 23, pp. 135–151, 1996.

[5] K. Fukunaga *Introduction to Statistical Pattern Recognition*, San Diego, CA: Academic, 1990.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Berlin, Germany: Springer-Verlag, 2009.

[7] E. Alpaydin, *Introduction to Machine Learning*, Cambridge, MA: MIT Press, 2004.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. Orlando, FL: Academic, 2006.

[9] C. T. Lin and C. S. George Lee, *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*, Englewood Cliffs, NJ: Prentice-Hall, 1996.

[10] R. B. Ska, P. J. van der Veen, and U. Kaymak, "Improved covariance estimation for Gustafson–Kessel clustering," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2002, pp. 1081–1085.

[11] B. S. Sebastiano and M. Gabriele, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, Feb. 2007.

[12] I. Jlliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[13] V. Zubko, Y. J. Kaufman, R. I. Burg, and J. V. Martins, "Principal component analysis of remote sensing of aerosols over oceans," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 730–745, Mar. 2007.

[14] A Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[15] B. C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.

[16] B. C. Kuo, D. A. Landgrebe, L. W. Ko, and C. H. Pai, "Regularized feature extractions for hyperspectral data classification," in *Proc. IGARSS*, Toulouse, France, Jul. 21–25, 2003, pp. 1767–1769.

[17] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Programming*, vol. 107, no. 3, pp. 391–408, 2006.

[18] R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Comput. J.*, vol. 6, pp. 163–168, 1963.

[19] D. Goldfarb, "A family of variable metric updates derived by variational means," *Math. Comput.*, vol. 24, pp. 23–26, 1970.

[20] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd ed. New York, NY: Springer, 2009.

[21] L. I. Kuncheva, and D. P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1798–1808, Nov. 2006.

[22] L. I. Kuncheva. (2010). [Online]. Available: http://www.bangor.ac.uk/∼mas00a/activities/artificial_data.htm.

[23] C. L. Blake and C. J. Merz (1998), "UCI repository of machine learning databases,", [Online]. Available: http://www.ics.uci.edu/∼mlearn MLRepository.html.

[24] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[25] N. Pal, K. Pal, J. Keller, and J. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 517–530, Aug. 2005.

[26] K. L. Wu, J. Yu, and M. S. Yang, "A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests," *Pattern Recognit. Lett.*, vol. 26, no. 5, pp. 639–652, 2005.

[27] S. Chatzis and T. Varvarigou, "Robust fuzzy clustering using mixtures of student's-t distributions," *Pattern Recognit. Lett.*, vol. 29, no. 13, pp. 1901–1905, Oct. 2008.

[28] S. Chatzis and T. Varvarigou, "Factor analysis latent subspace modeling and robust fuzzy clustering using t-distributions," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 505–517, Jun. 2009.

[29] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with fuzzy covariance matrix," in *Proc. IEEE CDC*, San Diego, CA, 1979, pp. 761–766.

[30] J. C. Bezdek and J. C. Dunn, "Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions," *IEEE Trans. Comput.*, vol. C-24, no. 8, pp. 835–838, Aug. 1975.

[31] N. R. Pal, K. Pal, and J. C. Bezdek, "A mixed c-means clustering model," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 1997, pp. 11–21.

[32] D. Tran and M. Wagner, "A robust clustering approach to fuzzy Gaussian mixture models for speaker identification," in *Proc. 3rd Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, 1999, pp. 337–340.

[33] P. F. Hsieh, D. S. Wang, and C. W. Hsu, "A linear feature extraction for multi-class classification problems based on class mean and covariance discriminant information," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 28, no. 2, pp. 223–235, Feb. 2006.

[34] F. Murtaugh, J. L. Starck, and M. W. Berry, "Overcoming the curse of dimensionality in clustering by means of the wavelet transform," *Comput. J.*, vol. 43, pp. 107–120, 2000.

[35] C. K. Wikle and N. Cressie, "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, vol. 86, pp. 815–829, 1999.
[36] E. Alphonse and S. Matwin, "Filtering multi-instance problems to reduce dimensionality in relational learning," *J. Intell.*, vol. 22, no. 1, pp. 23–40, 2004.
[37] R. Kohavi and D. Sommerfield, "Feature subset selection using the wrapper method: Overfitting and dynamic search space topology," in *Proc. First Int. Conf. Knowl. Discov. Data Mining*, Montreal, QC, Canada, 1995, pp. 192–197.
[38] R. Kohavi and G. H. John, "Wrappers for feature subset selection," in *Proc. Artif. Intell.*, 1997, pp. 273–324.
[39] S. Marinai, E. Marino, and G. Soda, "Embedded map projection for dimensionality reduction-based similarity search," in *Proc. Joint IAPR Int. Workshop Struct., Syntactic, Stat. Pattern Recognit.*, 2008, pp. 582–591.
[40] S. Marinai, E. Marino, and G. Soda, "Nonlinear embedded map projection for dimensionality reduction," *Lect. Notes Comput. Sci.*, vol. 5716, pp. 219–228, 2009.
[41] M. Dash and H. Liu, "Hybrid search of feature subsets," in *Proc. Pac. Rim 5th Int. Conf. Artif. Intell.*, 1998, Singapore, pp. 22–27.
[42] Z. Yin, Y. Tang, F. Sun, and Z. Sun, "Fuzzy clustering with novel separable criterion," *Tsinghua Sci. Tech.*, vol. 11, no. 1, pp. 50–53, Feb. 2006.
[43] S. J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in kernel Fisher discriminant analysis," in *Proc. Int. Conf. Mach. Learn.*, Pittsburgh, PA, 2006, pp. 465–472.
[44] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 460–474, Mar. 2005.
[45] A. Nasser and D. Hamad, "K-means clustering algorithm in projected spaces," in *Proc. 9th Int. Conf. Inf. Fusion*, 2006, pp. 1–6.
[46] K. Weike, P. Azad, and R. Dillmann, "Fast and robust feature-based recognition of multiple objects," in *Proc. 6th IEEE-RAS Int. Conf. Humanoid Robots*, 2006, pp. 264–269.
[47] K. Honda, A. Notsu, and H. Ichihashi, "Fuzzy PCA-guided robust k-means clustering," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 1, pp. 67–79, Feb. 2010.
[48] G. McLachlan, R. Bean, and L. B.-T. Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution," *Comp. Stat. Data Anal.*, vol. 51, no. 11, pp. 5327–5338, 2007.
[49] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 508–516, Aug. 2005.
[50] Z. Ghahramani and G. Hinton, "The EM algorithm for mixtures of factor analyzers," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRGTR-96-1, 1997.
[51] S. Chatzis. (2010). [Online]. Available: http://web.mac.com/soteri0s/Sotirios_Chatzis/Software.html
[52] S. T. John and C. Nello *Kernel Methods for Pattern Analysis*, Cambridge, U.K., Cambridge Univ. Press, 2004.
[53] P. S. Szczepaniak, P. J. G. Lisboa, and J. Kacprzyk, *Fuzzy Systems in Medicine*. Heidelberg, Germany/New York: Physica-Verlag, 2000.

Cheng-Hsuan Li received the B.S. and M.S. degrees from the National Chung Hsing University, Taichung, Taiwan, in 2001 and 2003, respectively. He is currently working toward the Ph.D. degree with the Institute of Electrical Control Engineering, National Chiao Tung University, Hsinchu, Taiwan.

He is also currently an Executive Assistant with the Graduate Institute of Educational Measurement and Statistics, National Taichung University. His research interests include pattern recognition and machine learning.

Bor-Chen Kuo (S'01–M'02) received the B.S. degree in mathematics education and the M.S. degree in statistics and assessment from National Taichung Teachers College, Taichung, Taiwan, in 1993 and 1996, respectively, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 2001.

From 2002 to 2005, he was an Associate Professor with National Taichung University of Education, where he is currently a Professor and the Director of the Graduate Institute of Educational Measurement and Statistics. His research interests include pattern recognition, feature extraction and classification of high-dimensional data, computerized adaptive testing, and nonparametric functional estimation.

Dr. Kuo is a member of the Chinese Association of Psychological Testing Administrative Committee and the Editor of the *Journal of Educational Measurement and Statistics*, Taiwan. He was the recipient of an Outstanding and Excellence Research Award from the R.O.C. Education and Research Society in 2009.

Chin-Teng Lin (F'05) received the B.S. degree from the National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1986 and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1992.

He is currently the Chair Professor of electrical and computer engineering with NCTU. He has authored or coauthored more than 120 journal papers in the areas of neural networks, fuzzy systems, multimedia hardware/software, and cognitive neuroengineering, including approximately 74 IEEE journal papers. His research interests include biologically inspired information systems.