Neurocomputing 103 (2013) 29–42

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Regularized soft *K*-means for discriminant analysis

Xuesong Yin [a,b,*], Songcan Chen [a], Enliang Hu [a]

[a] Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, China
[b] Department of Computer Science and Technology, Zhejiang Radio and TV University, China

ARTICLE INFO

Article history:
Received 4 August 2011
Received in revised form
15 May 2012
Accepted 20 August 2012
Communicated by J. Tin-Yau Kwok
Available online 8 October 2012

Keywords:
Soft clustering
Discriminant analysis
Data clustering
Dimensionality reduction

ABSTRACT

Traditionally unsupervised dimensionality reduction methods may not necessarily improve the separability of the data resided in different clusters due to ignorance of the inherent relationship between subspace selection and clustering. It is known that soft clustering using fuzzy c-means or its variants can provide a better and more meaningful data partition than hard clustering, which motivates us to develop a novel entropy regularized soft *K*-means algorithm for discriminant analysis (ResKmeans) in this paper. ResKmeans performs soft clustering and subspace selection simultaneously and thus gives rise to a generalized linear discriminant analysis (GELDA) which captures both the intra-cluster compactness and the inter-cluster separability. Furthermore, we clarify both the relationship between GELDA and conventional LDA and the inherent relationship between subspace selection and soft clustering. Experimental results on real-world data sets show ResKmeans is superior to other popular clustering algorithms.

1. Introduction

It is well known that data clustering plays an indispensable role in understanding and uncovering the structure of unlabeled data. To this end, a clustering algorithm should make similar instances group into the same cluster and dissimilar instances into different clusters with good separability to reflect the inherent structures in the data. Undoubtedly, a good clustering is crucial to many real-world applications especially for high-dimensional data such as digital images, financial time series and gene expression microarrays. Unfortunately, many known algorithms tend to work ineffectively in the high-dimensional space because the distance between every pair of instances is almost the same for a wide variety of data distributions and distance functions [1]. In fact, since the high-dimensional data often reside on or near an underlying low-dimensional manifold, one natural way to overcome the above limitation is to apply dimensionality reduction techniques to effectively seek a low-dimensional representation of the data for clustering. The techniques for unsupervised dimensionality reduction such as ISOMAP [2], constrained Laplacian Eigenmap [3], principal component analysis (PCA) [4] and its variant [5], locally linear embedding (LLE) [6], locality preserving projections [7] and sample-dependent graph construction [8], as a pre-processing step before clustering, may be difficult in improving the separability among the clusters for better characterizing the structures in data [9]. In order to alleviate this difficulty, one natural way is to borrow the idea from supervised dimensionality reduction such as linear discriminant analysis (LDA) [10–13] to separate the data by utilizing the obtained cluster labels. Following such a line, recently, researchers have proposed several iterative subspace selection and clustering algorithms to promote their own performances mutually, consequently forming so-called discriminative clustering. Currently, there are two major approaches of discriminative clustering: the LDA-guided approach and the distance metric learning-based approach. The former combines LDA and *K*-means clustering into a coherent framework to select the most discriminative subspace [1,14,15]. Specifically, this class of approaches uses *K*-means clustering to generate the cluster labels and then applies LDA to perform subspace selection. The latter performs clustering and distance metric learning simultaneously. Such approaches show that the joint clustering and distance metric learning can be formulated as a trace maximization problem, which can be solved via an alternating iterative procedure as done in the EM framework [9,16]. In practice, the above algorithms apply the *K*-means algorithm to cluster the instances in low-dimensional discriminative space where each instance only belongs to one single cluster, thus called "hard" clustering. Although such hard clustering is a well-posed problem, the crisp solution may be characterized by the extremal points [17,18]. On the other hand, soft clustering using fuzzy c-means and its variants have been shown to be capable of providing a better and more meaningful data partition than hard clustering approaches [19]. Moreover, hard clustering algorithms often fail to adequately reflect the real-world data description because of the uncertain nature of their memberships [20,21].

* Corresponding author at: Zhejiang Radio and TV University, Department of Computer Science and Technology, No. 6 Zhenhua Road, Hangzhou, Zhejiang Province 310030, China. Tel.: +86 571 88086839; fax: +86 571 89983165.
E-mail address: yinxs@nuaa.edu.cn (X. Yin).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.021

In this paper, we are more interested in how to uncover the structure in real-world data as faithfully as possible and improve the clustering accuracy for the high-dimensional data. To this end, we follow the first approach above and propose a new regularized soft K-means algorithm for discriminant analysis, called ResKmeans. More specifically, through relaxing the membership degree of the data from hard binary values of {0,1} to the soft interval [0,1] and subsequently adding an entropic regularization term to the objective function of K-means [22], thus a regularized soft K-means algorithm is developed and then applied to cluster the data. However, different from existing typical algorithms [1,13,15], we specially develop a generalized linear discriminant analysis approach (GELDA) for adaptively selecting the most discriminative subspace based on the obtained membership degrees rather than the cluster labels. With such a new GELDA, the new data representations can be generated by projecting all the data points onto the most discriminative subspace, and further, are again input to the regularized soft K-means algorithm to find the appropriate cluster assignments. In a word, the procedure both for finding the most discriminative subspace using GELDA and for clustering the data points will alternatively perform until the appropriate cluster assignments are obtained.

Fig. 1 illustrates the data distributions based on the new representations in the projected subspaces. The data points used in this illustration are sampled from the News-Different data set (described later in Section 4.1) which belong to the three clusters as labeled in Fig. 1 by legends Δ, $\circ$, and $+$, respectively. Sub-figure (a) shows the original data distribution projected into a 2D space generated by PCA. We see from it clearly that many data points of the cluster $\circ$ overlap heavily with data points of the clusters Δ and $+$, and are difficult to be separated. Sub-figures (b) and (c) illustrate the results generated by LLE and PCACL, respectively. Sub-figure (d) shows the data distribution in the newly-projected subspace, in which clustering and distance metric learning are performed simultaneously. (e) and (f) illustrate the results generated by LDA-KM and DisKmeans, respectively. (g) illustrates the projected data distribution based on the new representations generated by the proposed method, which are helpful in separating the data points in the cluster $\circ$ from those in the other two clusters. Visually, the proposed algorithm can better improve the separability of the cluster in the projected subspace in comparison to the other representative methods. In what follows, we summarize favorable and attractive characteristics of the proposed algorithm:

(1) we propose a general discriminative clustering framework whose novelty lies in that it integrates GELDA and regularized soft K-means. Under the framework, previous work [1] becomes its special case when the membership degree in soft K-means is constrained to the binary values;
(2) maximum entropy is introduced as a regularized term such that ResKmeans obtains more effective partitions with the help of GELDA than other similar works, which gains some new insights into the description of the real-world data;
(3) we show that GELDA and soft K-means clustering optimize the same optimization criterion to achieve both minimization of the within-cluster scatter and maximization of the between-cluster scatter;
(4) the formulation of GELDA generalizes the famous LDA;
(5) last, ResKmeans leads to a natural generalization for existing similar works and can also accommodate other off-the-shelf soft clustering algorithms such as mixture of Gaussians, fuzzy c-means and so on.

The rest of this paper is organized as follows. In Section 2, we briefly review some related work on LDA subspace selection and K-means clustering. Then we formulate the proposed general discriminative clustering framework and derive the ResKmeans algorithm in Section 3. We report experimental results in Section 4 and finally conclude this paper in Section 5.

2. Related work

Before we describe related work and the proposed method, we summarize the main notations used in this paper in Table 1.

2.1. Discriminant analysis and K-means clustering

Let X denote a data set with n instances, $\{x_i\}_{i=1}^n \in \Re^D$. For simplicity, the data are centered, i.e., $m=\sum_{i=1}^n x_i/n=0$. Denote $X=[x_1,\ldots,x_n]$ as the data matrix whose i-th column is given by x_i. We use $trace(M)$ to denote the trace of matrix M. The matrix $H=\{0,1\} \in \Re^{n\times K}$is the cluster indicator: $H_{ik}=1$ if x_i belongs to the cluster C_k, and $H_{ik}=0$ otherwise. K-means clustering is to minimize the following objective function:

$$J_K = \sum_{k=1}^{K} \sum\nolimits_{x_i \in C_k} \|x_i - m_k\|^2, \tag{1}$$

where m_k denotes the centroid of the cluster C_k.

Ding et al. [1] proposed an adaptive dimension reduction algorithm, called LDA-Km. Specifically, LDA-Km uses LDA and K-means clustering to iteratively perform subspace selection and clustering. After obtaining the indicator H by Eq. (1), LDA-Km respectively defines the between-cluster scatter and within-cluster scatter matrices as follows:

$$S_w = \left(X - MH^T\right)\left(X - MH^T\right)^T, \tag{2}$$

$$S_b = MH^T HM^T, \tag{3}$$

LDA-Km optimizes the following objective function to obtain the projection matrix W that maps the high-dimensional data onto the low-dimensional space:

$$\max_{W,H} \frac{trace\left(W^T S_b W\right)}{trace\left(W^T S_w W\right)}. \tag{4}$$

Thus, LDA-Km uses K-means clustering to generate cluster labels and uses LDA to do subspace selection until an (local) optimal solution is obtained. Similar to LDA-Km, Torre et al. [14] proposed the discriminative cluster analysis (DCA) algorithm and optimized the following objective function:

$$\max_{W,H} \frac{trace\left(W^T S_b W\right)}{trace\left(W^T S_t W\right)}. \tag{5}$$

In practice, the denominator of Eq. (5) is $W^T S_t W$ rather than $W^T S_w W$ which is crucial step in LDA, such that Eq. (5) is not a full LDA in contrast to LDA-Km.

2.2. Discriminative K-means for clustering

Ye et al. [15] proposed the discriminative K-means (DisKmeans) algorithm for simultaneous LDA subspace selection and clustering. First of all, the cluster indicator is defined as follows:

$H = h_{ij_{n\times K}}$ where $h_{ij} = 1$, iff $x_i \in C_j$

DisKmeans defines the weighted cluster indicator L with H as follows:

$$L = [L_1,\ldots,L_K] = H\left(H^T H\right)^{-1/2}. \tag{6}$$

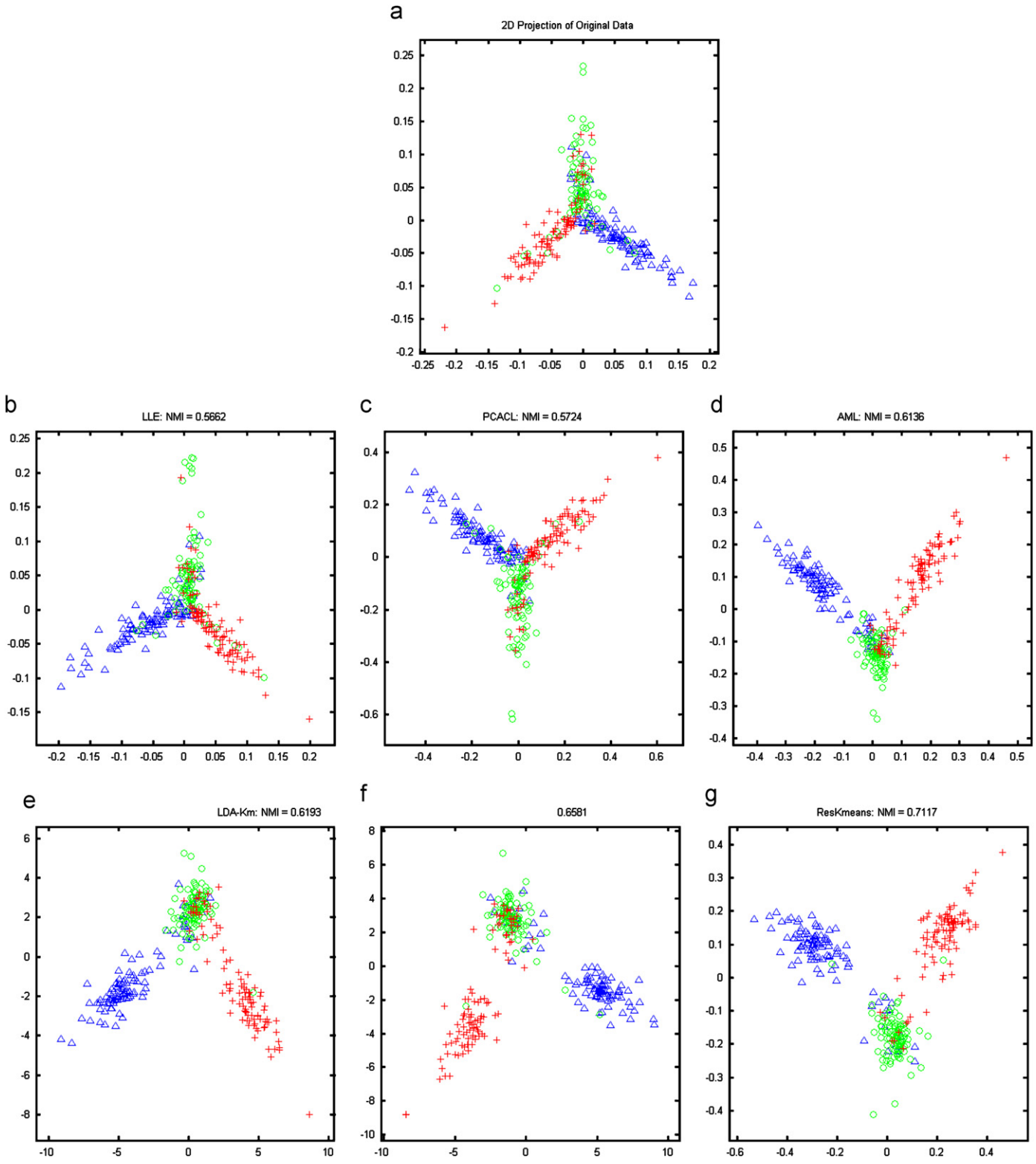

Fig. 1. An example illustrating the new representations in the projected subspaces. (a) shows the original data distribution, projected to the space spanned by its two principal components; (b)-(g) show the corresponding results obtained by existing representative methods and our method respectively. (a) Input Data, (b) LLE, (c) PCACL, (d) AML, (e) LDA-Km, (f) DisKmeans and (g) ResKmeans.

It follows that the j-th column of L is obtained by

$$L_j = \left(0,\ldots,0,\overbrace{1,\ldots,1}^{n_j},0,\ldots,0\right)^T / n_j^{1/2}$$

Hence, the corresponding within-cluster scatter, between-cluster scatter, and total scatter matrices are defined as follows:

$$S_w = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - m_k)(x_i - m_k)^T, \tag{7}$$

$$S_b = XLL^TX^T,\ S_t = XX^T, \tag{8}$$

Since the estimation of the total scatter matrix is often unreliable for the high-dimensional data, DisKmeans applies the regularization technique to improve its estimation as follows:

$$\tilde{S}_t = S_t + \lambda I_D = XX^T + \lambda I_D. \tag{9}$$

The above Eq. (5) can thus be rewritten as

$$f(L,W) = trace\left(\left(W^T\left(XX^T + \lambda I_D\right)W\right)^{-1}W^TXLL^TX^TW\right). \tag{10}$$

The optimal projection matrix W can be expressed as $W=XP$ for$P \in \Re^{n\times d}$. Further, solving the optimal W is transformed to solve the matrix P which can be obtained by maximizing the following objective function:

$$f(L,W) = trace\left(\left(P^T(GG+\lambda G)P\right)^{-1}P^TGLL^TGP\right), \tag{11}$$

Where $G=X^TX$ is called as the Gram matrix. It was shown that the optimal solution of the matrix L can be obtained by solving the following maximization problem:

$$L^* = \underset{L}{\operatorname{argmax}}\ trace\left(L^T\left(I_n - \left(I_n + \frac{1}{\lambda}G\right)^{-1}\right)L\right) \tag{12}$$

Different from LDA-Km and DCA, DisKmeans can be dramatically simplified by removing the iteration between subspace selection and clustering. In short, the above methods generally adopt hard K-means to cluster the data in the projected space. However, one of the shortcomings of hard clustering methods is not to adequately take the fuzzy nature of the real-world data memberships into account. In this paper, we aim to devise a general framework of discriminant analysis by adopting regularized soft K-means to realize both subspace selection and clustering. With the framework, we develop the ResKmeans algorithm to overcome the shortcoming above and further improve the clustering performance over state-of-the-art approaches.

3. Regularized soft K-means for discriminant analysis

3.1. Regularized soft K-means

Instead of hard K-means that each instance belongs to one and only one cluster, soft K-means algorithms [23–25] group each instance into one cluster in terms of the membership degree by optimizing the following objective function:

$$\min J_s = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\|x_i - m_k^r\|^2 \tag{13}$$

$$\text{subject to } \sum_{k=1}^{K} u_{ik} = 1,\ u_{ik} \geq 0 \tag{14}$$

where u_{ik} is the membership degree of x_i belonging to the cluster C_k whose centroid is m_k^r. Let $U=[u_{ik}]\in[0,1]^{n\times K}$. In order to avoid overfitting in the clustering process and obtain a probability assignment for the instance, entropy, as the regularization term [17,26], is introduced to the objective function of soft K-means, thus the regularized objective is formed as

$$\min J^r = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\|x_i - m_k^r\|^2 + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\ln u_{ik} \tag{15}$$

$$\text{subject to } \sum_{k=1}^{K} u_{ik} = 1,\ u_{ik} \geq 0. \tag{16}$$

Table 1
Summary of notations.

Symbols	Description
n	Number of instances
K	Number of pre-specified clusters
$X=\{x_i\}_{i=1}^n$	Set of n unlabeled instances
H	Cluster indicator
L	Weighted cluster indicator
W	Projection matrix
U	Membership degree
S_w^r	(soft) Within-cluster scatter matrix
S_b^r	(soft) Between-cluster scatter matrix
S_t^r	(soft) total scatter matrix
η	Regularization parameter

To solve u_{ik} which satisfies constraints (16), we further define a Lagrangian function of (15) as follows:

$$L(u_{ik},\lambda) = J^r - \lambda\left(\sum_{k=1}^{K} u_{ik} - 1\right), \tag{17}$$

where λ is a Lagrangian multiplier. Now taking the derivative of $L(u_{ik},\lambda)$with respect to $u_{ik}$and setting it to zero, we get

$$\frac{\partial L(u_{ik},\lambda)}{\partial u_{ik}} = \|x_i - m_k^r\|^2 + \eta(\ln u_{ik} + 1) - \lambda = 0.$$

Solving the above equation, we have

$$u_{ik} = \frac{\exp\left(-\|x_i - m_k^r\|^2/\eta\right)}{\sum_{k=1}^{K}\exp\left(-\|x_i - m_k^r\|^2/\eta\right)}. \tag{18}$$

Similarly, we also obtain the corresponding clustering centroid:

$$m_k^r = \sum_{i=1}^{n} u_{ik}x_i \Big/ \sum_{i=1}^{n} u_{ik}. \tag{19}$$

Therefore, with the regularization term of maximum entropy, the membership assignments obtained by soft K-means clustering not only satisfy constraints (16) to have a probabilistic interpretation but also can reflect the inherent fuzzy nature of the data. Particularly, if the assignment is just taken as 0 or 1, soft K-means clustering reduces to hard K-means clustering.

3.2. Generalized linear discriminant analysis

Based on the membership U obtained in Section 3.1, in this subsection, we will develop a GEneralized Linear Discriminant Analysis method (GELDA). Like LDA, one key is to require defining the within-cluster scatter, between-cluster scatter and total scatter matrices for establishment of a discriminant criterion, but unlike LDA which uses the supervised class labels, now just the instances' membership degrees from soft clustering are available and then will be used to respectively define the so-needed (soft) within-cluster scatter matrix S_w^r, (soft) between-cluster scatter matrix S_b^r and (soft) total scatter matrix S_t^r for GELDA as in (20–22)

$$S_w^r = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i - m_k^r)(x_i - m_k^r)^T, \tag{20}$$

$$S_b^r = \sum_{k=1}^{K} n_k^r(m_k^r - m^r)(m_k^r - m^r)^T, \tag{21}$$

$$S_t^r = \sum_{i=1}^{n}(x_i - m^r)(x_i - m^r)^T, \tag{22}$$

where $n_k^r = \sum_{i=1}^{n} u_{ik}$ is the (fuzzy or soft) number of the instances involved in the cluster C_k. m^r is the (fuzzy) global center of the data, which can be expressed as

$$m^r = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}x_i}{\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}} \tag{23}$$

since $\sum_{k=1}^{K} u_{ik} = 1$, we have $\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik} = n$

and

$$m^r = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}x_i = \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{k=1}^{K} u_{ik} = \frac{1}{n}\sum_{i=1}^{n} x_i = m.$$

Thus, m^r is equal to m (the global or sample center of the data in hard clustering). Especially, both m^r and m are zero when the data are centralized.

Further, we obtain a more general formulation for discriminant criterion and two insights: The first insight is that as a generalization of LDA, GELDA can make effective use of the fuzzy nature of the real-world data to seek the most discriminative subspace. The relationship between GELDA and LDA is summarized in the following proposition 1.

Proposition 1. *S_t^r in GELDA and S_t in LDA are identical*, i.e.,

$$S_t^r = S_t. \tag{24}$$

Clearly, since m^r is equal to m, Proposition 1 holds.

Accordingly, since the value of u_{ik} lies in the interval [0,1] and $n_k^r = \sum_{i=1}^{n} u_{ik}$, usually $n_k^r \neq n_k$ unless each cluster is pure, i.e., all the instances in the cluster share a common class label. m_k^r and m_k are respectively given by

$$m_k^r = \sum_{i=1}^{n} u_{ik}x_i \Big/ \sum_{i=1}^{n} u_{ik} = \sum_{i=1}^{n} u_{ik}x_i / n_k^r,$$

$$m_k = \sum_{i=1}^{n_k} x_i / n_k.$$

Thus generally, $m_k^r \neq m_k$. Recalling Eqs. (3) and (21), in most cases, also $S_b^r \neq S_b$ and $S_w^r \neq S_w$ from Eqs. (7) and (20).

The second insight from the formulations of GELDA is that trace(S_w^r) and trace(S_b^r) capture the intra-cluster compactness and the inter-cluster separability, respectively. Moreover, it can be shown that $S_t^r = S_w^r + S_b^r$ (See Appendix later). Thus, we can apply GELDA to find the directions along which instances of different clusters can be far from each other while those of the same cluster are close to each other. Such an optimal projection can be obtained by optimizing the following objective function:

$$\max \frac{trace\left(W^T S_b^r W\right)}{trace\left(W^T S_w^r W\right)}. \tag{25}$$

Similar to LDA-Km, GELDA and soft K-means clustering optimize the same optimization criterion, which can be summarized in the following proposition 2.

Proposition 2. *Suppose S_w^r, S_b^r, S_t^r to be the within-cluster scatter, between-cluster scatter and total scatter matrices in GELDA respectively, and J_s to be the objective function of soft K-means clustering. Then GELDA and soft K-means clustering have the same optimization criterion.*

Proof: The goal of soft *K*-means clustering is to optimize the following objective function under the constraints $\sum_{k=1}^{K} u_{ik} = 1$ and $u_{ik} \geq 0$:

$$\min J_s = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\|x_i - m_k^r\|^2$$

Since $S_t^r = S_w^r + S_b^r$, the above objective function can be written as

$$\begin{aligned} J_s &= trace\left(\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i - m_k^r)(x_i - m_k^r)^T\right) \\ &= trace\left(S_w^r\right) = trace\left(S_t^r - S_b^r\right) \end{aligned} \tag{26}$$

Therefore, soft *K*-means clustering minimizes the within-cluster scatter, equivalently, maximizes the between-cluster scatter since $S_t^r = S_t$ is a constant matrix for the given data.

Additionally, S_w^r and S_b^r are respectively derived by the membership degree $\{u_{ik}\}$ so that the optimal projection W^* of GELDA can be determined by optimizing Eq. (25). As a result, we minimize the within-cluster scatter while maximizing the between-cluster scatter. Thus, GELDA and soft *K*-means clustering have the same optimization criterion. This completes the proof of this proposition. □

Based on the theoretical analysis above, if we get the projection matrix W and fix it, the membership degree matrix U can be computed by minimizing the following entropy-regularized objective function subject to (16):

$$\begin{aligned} f(U) &= trace\left(W^T S_w^r W) + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\ln u_{ik}\right) \\ &= trace\left(W^T \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i - m_k^r)(x_i - m_k^r)^T W\right) + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\ln u_{ik} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\|W^T x_i - W^T m_k^r\|^2 + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\ln u_{ik} \end{aligned} \tag{27}$$

Obviously, the above formulation is exactly a regularized soft *K*-means in the projected subspace $Y = W^T X$. Analogously, when U is fixed, we can compute W using GELDA. Essentially, the optimization is alternatively carried out by fixing one of two components (W and U) and then optimizing the other. The corresponding algorithm, called ResKmeans is presented in Algorithm 1 below.

Algorithm 1: ResKmeans.

Input: Dataset X, cluster number K;
Output: Membership degree U, projection W.
Step 1: Set dimension $d = K$-1;
Step 2: Compute PCA on X to obtain the initial W;
Step 3: Perform regularized soft *K*-means: compute matrix U by Eq. (18) in the W-transformed space;
Step 4: Perform GLDA: get S_w^r, S_b^r and S_t^r with U by Eqs. (20–22) to compute W by Eq. (25);
Step 5: Go to step 3 until convergence;
Step 6: Return U, W.

After the iteration for finding W and U, we can obtain the new projection space and K partitions of all the data. Concretely, if u_{ik} is greater than the other membership degrees in the i-th row, x_i belongs to the k-th cluster C_k. We thus group n data points into K clusters according to obtained U.

Next, we discuss the convergence of ResKmeans. Our goal is to solve (U, M^r) and W jointly by optimizing the following problem:[1]

$$\min J(W,U,M^r) = trace\left(W^T S_w^r W\right) + \eta\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}\ln u_{ik} \tag{28}$$

[1] If S_t is singular, we can also apply the regularization technique as (9) to improve its estimation.

$$\text{subject to } trace\left(W^T S_t W\right) = 1, \tag{29}$$

$$\sum_{k=1}^{K} u_{ik} = 1,\ u_{ik} \geq 0, \tag{30}$$

Where

$M^r = [m_1^r, \ldots, m_K^r]$

Due to the nonconvexity of the above optimization problem with respect to (U, M^r, W), directly solving them simultaneously is difficult. Instead we adopt the alternating optimization technique as in fuzzy K-means [23] to optimize them. In this way, for a given W, the objective function (28) which is now just subject to the constraints (30) and independent on the constraint (29), naturally boils down to our familiar soft K-means clustering in the W-transformed space. Thus according to [23], (U, M^r) can alternatively be solved between U and M^r and finally converge to some local optimum in finite time iterations. Next, for a given (U, M^r), optimizing (28) with respect to W, just dependent on the constraint (29), is equivalent to optimizing (25). Though this problem is not convex, we can still obtain its approximate analytic solution by solving the same generalized eigenvalue equation as LDA. This process is alternatively repeated till an approximate solution or some local optimum of (U, M^r, W) is reached. Again due to nonconvexity of the constraint (29), the theoretical convergence of ResKmeans is generally difficult to be proved, which is exactly the same theoretical problem as Ding et al. confronted in their LDA-Km [1]. Consequently, following [1], we also have to resort to an empirical means to examine the convergence of ResKmeans. Fortunately, in the following experiments, we find that ResKmeans typically converges within 10 iterations on all the data sets used here, which is likewise consistent with Ding et al.'s findings.

Finally, ResKmeans consists of two parts: GELDA and soft K-means clustering. It is equivalent to t (GELDA+soft K-means clustering), where t is the iteration number of the algorithm to converge. Therefore, the time complexity of ResKmeans is $O(Dnt)$ for soft K-means clustering and $O(D^2nt)$ for GELDA where D is the dimension of data and n is the number of the data points.

4. Experiments

In this section, we present an experimental study to evaluate the ResKmeans algorithm in comparison to several representative algorithms. Particularly, we aim to address the following three questions in our study:

(1) As a general discriminative clustering framework, is the proposed algorithm able to effectively integrate GELDA subspace selection and regularized soft K-means?
(2) How effective is the proposed discriminative framework compared with the other discriminative clustering algorithms?
(3) Can the imposed regularization make the iterative subspace selection and clustering process be prone to stable?

4.1. Data sets

We compare all the algorithms on the fifteen data sets, including Digits, Iris, Letter (Letter (a–d)), Protein, Segment, Wine, Zoo from UCI Machine Learning Repository, three image data sets: USPS handwritten data, Yale face, ORL face, and three document data sets: News-Similar, News-Same and News-Different from 20 Newsgroups. The statistics of all the data sets are summarized in Table 2.

4.2. Clustering evaluation

For comprehensive evaluation on the clustering performance, in the following experiments, we adopt two popular measures frequently used in both clustering algorithm and dimensionality reduction, i.e. Accuracy [1,14–16,27] and normalized mutual information [14,15,27,28]. The accuracy (ACC) is defined as follows:

$$\text{ACC} = \frac{1}{n}\sum_{i=1}^{n} \delta(s_i, map(r_i))$$

Where r_i and s_i are respectively the obtained cluster label and the ground-truth label of instance x_i, n is the number of the instances, $\delta(x,y)$ equals 1 if $x=y$ and 0 otherwise, and $map(r_i)$ denotes a permutation mapping from the cluster label r_i to the ground-truth label from the data set.

The normalized mutual information (NMI) is defined as

$$\text{NMI} = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{31}$$

where X and Y denote the random variables of the cluster memberships from the ground truth and the output of clustering algorithm, respectively, and $I(X,Y)$ measures the mutual information between X and Y. $H(X)$ and $H(Y)$ are the entropies of X and Y respectively.

4.3. Comparison with the representative work

To validate that the proposed discriminative framework is capable of effectively integrating subspace selection and regularized soft K-means, other representative algorithms are used in our study. They are

(1) PCA+K-means, in which PCA [4] is applied to select the subspace and the K-means algorithm is used to cluster the projected data;
(2) LLE+K-means, in which LLE [6] is applied to select the subspace and K-means is used to cluster the projected data;
(3) LDA-Km [1], in which LDA and K-means are combined into a coherent framework to select the subspace and perform clustering simultaneously;
(4) AML [9], an unsupervised adaptive metric learning algorithm which performs clustering and distance metric learning simultaneously;
(5) DisKmeans [15], a simultaneous LDA subspace selection and clustering algorithm which is a simplified version of LDA-Km;
(6) PCACL [5], in which the classical competitive learning is extended by performing a principal components analysis;
(7) brFCM [24], an accurate fuzzy clustering algorithm which is mainly used to cluster the low-dimensional data.

We implement our ResKmeans and the above seven algorithms in the MATLAB environment. All experiments were conducted on a PENTIUM DUAL 1.6 G PC with 1.0 GB RAM. For each data set, we repeated experiments for 20 trials and tested the performance on the whole data set. Also, all the data is normalized and the average results from 20 trials are reported.

We use PCA+K-means as the baseline algorithm for comparison. Since brFCM is mainly used to cluster the low-dimensional data, PCA is applied to reduce the dimension of the data before it clusters the high-dimensional data. Additionally, the regularization parameter η is determined by searching in $\{10^{-3}, 10^{-2}, 10^{-1}\}$. Like [1,9,14–16,29], for all the (compared) eight algorithms, the cluster number K and the reduced dimensionality of the data are simply set to the

actual number of classes and K-1, respectively. Experimental results are shown in terms of the low-dimensional and the high-dimensional data sets respectively.

Table 2
Summary of the test datasets used in our experiment.

Dataset	Instance	Dimension	Class
ORL	400	4096	40
Yale	165	2048	15
News-Different	300	1000	3
News-Same	295	1000	3
News-Similar	288	1000	3
USPS3568	3084	256	4
USPS	3000	256	10
Digits(0–5)	332	16	6
Digits	7479	16	10
Iris	150	4	3
Letter(a–d)	3096	16	4
Protein	116	20	6
Segment	1980	20	6
Wine	178	13	3
Zoo	101	18	7

(1) **Results on the low-dimensional data sets**. Table 3 shows the clustering results of all the algorithms by the ACC and NMI measures on the low-dimensional data sets. For simplicity, we use the same initial inputs for these eight approaches. Obviously, we can observe that ResKmeans outperforms all the compared methods on the five data sets: Digits(0–5), Digits, Protein, Zoo and Letter (a–d). DisKmeans performs best on Segment and Wine. Similar to LDA-Km, AML performs best just on one data set. On the whole, the clustering performance of ResKmeans is slightly better than those of the other seven clustering methods on the low-dimensional data sets. Interestingly, we see a common phenomenon that all the discriminative clustering algorithms including LDA-Km, AML, DisKmeans and ResKmeans always outperform the non-discriminant ones such as PCA+K-means and LLE+K-means.

(2) **Results on the high-dimensional data sets**. Table 4 shows the clustering accuracies of all the approaches on seven high-dimensional data sets. It is interesting to note that ResKmeans outperforms all compared methods on five data sets such as ORL, Yale, News-Different, News-Same and News-Similar. Also, on the rest data sets, ResKmeans can compare with the

Table 3
ACC and NMI results on the low-dimensional data sets (mean ± std-dev%).

Dataset	PCA	LLE	PCACL	brFCM	LDA-Km	AML	DisKmeans	ResKmeans
(a) ACC comparison on the low-dimensional data sets								
Digits (0–5)	84.94 ± 3.82	83.13 ± 3.91	85.12 ± 3.77	85.24 ± 3.64	87.05 ± 3.05	**87.95 ± 2.59**	86.24 ± 2.55	**87.95 ± 2.32**
Digits	77.63 ± 4.10	73.27 ± 4.25	76.39 ± 4.01	77.86 ± 3.98	80.00 ± 3.62	79.09 ± 3.76	78.12 ± 3.81	**80.36 ± 3.54**
Iris	88.67 ± 2.24	77.33 ± 2.52	89.29 ± 1.72	89.57 ± 1.84	**98.00 ± 0.70**	96.00 ± 0.96	95.13 ± 0.84	96.67 ± 0.79
Letter (a–d)	63.34 ± 6.76	63.78 ± 6.64	62.89 ± 6.38	63.60 ± 6.41	64.50 ± 6.11	66.21 ± 5.03	65.02 ± 5.26	**68.93 ± 5.45**
Protein	56.52 ± 4.41	57.39 ± 4.35	56.91 ± 4.05	57.44 ± 3.87	67.83 ± 3.47	61.74 ± 3.92	59.13 ± 4.36	**68.70 ± 3.21**
Segment	56.57 ± 5.56	57.34 ± 5.52	57.69 ± 5.30	58.09 ± 5.05	72.28 ± 5.26	75.92 ± 3.59	**76.37 ± 4.41**	73.39 ± 4.35
Wine	65.23 ± 3.63	65.87 ± 3.56	66.04 ± 3.03	66.30 ± 3.12	69.66 ± 3.20	71.66 ± 1.87	**72.83 ± 2.99**	69.66 ± 3.15
Zoo	76.24 ± 3.86	80.20 ± 3.40	77.63 ± 3.14	78.63 ± 3.09	85.32 ± 2.33	84.16 ± 3.23	85.18 ± 2.44	**86.14 ± 2.21**
(b) NMI comparison on the low-dimensional data sets								
Digits(0–5)	75.30 ± 3.26	73.38 ± 3.34	75.71 ± 3.17	75.89 ± 3.15	78.90 ± 2.61	**79.28 ± 2.24**	76.59 ± 2.39	78.72 ± 2.20
Digits	69.69 ± 3.80	68.68 ± 3.87	69.04 ± 3.65	69.23 ± 3.60	75.06 ± 3.22	74.99 ± 3.29	70.52 ± 3.40	**76.70 ± 3.14**
Iris	74.19 ± 1.71	65.73 ± 1.83	75.16 ± 1.22	75.20 ± 1.25	**89.42 ± 0.69**	86.42 ± 0.88	85.25 ± 0.81	88.51 ± 0.75
Letter(a–d)	50.63 ± 5.56	50.76 ± 5.50	49.80 ± 5.39	50.64 ± 5.27	50.86 ± 5.06	51.24 ± 4.47	51.71 ± 4.89	**52.23 ± 4.67**
Protein	42.66 ± 4.21	42.77 ± 4.13	42.37 ± 4.02	43.00 ± 3.94	52.73 ± 3.22	46.30 ± 3.85	44.17 ± 4.07	**53.35 ± 3.09**
Segment	54.24 ± 4.99	54.29 ± 4.94	55.06 ± 4.88	55.78 ± 4.82	64.25 ± 4.63	68.81 ± 3.55	**69.59 ± 4.20**	65.99 ± 4.15
Wine	52.88 ± 3.07	53.32 ± 3.01	54.12 ± 2.79	54.54 ± 2.80	61.30 ± 2.84	64.79 ± 1.98	**65.82 ± 2.77**	62.59 ± 2.81
Zoo	74.15 ± 3.63	79.56 ± 3.35	74.88 ± 2.83	75.63 ± 2.74	83.51 ± 2.07	80.35 ± 2.79	81.10 ± 2.10	**84.01 ± 1.95**

Table 4
ACC and NMI results on the high-dimensional data sets (mean ± std-dev%).

Dataset	PCA	LLE	PCACL	brFCM	LDA-Km	AML	DisKmeans	ResKmeans
(a) ACC comparison on the high-dimensional data sets								
ORL	75.00 ± 2.40	80.15 ± 2.12	75.12 ± 2.37	76.23 ± 2.31	83.00 ± 1.93	80.75 ± 2.24	81.25 ± 2.02	**85.50 ± 1.78**
Yale	72.31 ± 1.94	75.52 ± 1.71	73.64 ± 1.93	74.04 ± 1.80	76.36 ± 1.67	76.24 ± 1.62	77.88 ± 1.53	**81.64 ± 1.31**
News-Different	77.41 ± 2.13	79.33 ± 1.95	78.60 ± 2.10	80.05 ± 1.86	82.67 ± 1.25	82.67 ± 1.23	83.43 ± 1.21	**86.57 ± 0.95**
News-Same	41.27 ± 3.41	42.62 ± 3.32	41.03 ± 3.39	42.98 ± 3.25	45.87 ± 2.70	44.90 ± 2.97	46.62 ± 2.89	**53.99 ± 2.47**
News-Similar	52.21 ± 3.50	56.94 ± 3.39	52.54 ± 3.46	53.61 ± 3.43	59.38 ± 3.08	63.11 ± 2.65	62.42 ± 2.77	**65.97 ± 2.15**
USPS3568	71.02 ± 1.78	69.84 ± 2.11	69.44 ± 2.03	72.11 ± 1.58	74.22 ± 1.29	74.55 ± 1.23	**76.78 ± 1.01**	74.11 ± 1.36
USPS	72.03 ± 2.50	67.74 ± 2.89	72.39 ± 2.44	73.20 ± 2.31	78.10 ± 1.75	76.47 ± 2.00	**79.10 ± 1.68**	78.90 ± 1.73
(b) NMI comparison on the high-dimensional data sets								
ORL	83.95 ± 1.80	86.36 ± 1.66	84.22 ± 1.86	84.59 ± 1.73	88.87 ± 1.52	87.31 ± 1.54	88.67 ± 1.41	**89.23 ± 1.39**
Yale	70.45 ± 2.92	71.35 ± 3.01	70.51 ± 2.91	71.02 ± 2.70	72.65 ± 2.54	72.33 ± 2.54	75.83 ± 2.05	**78.05 ± 1.81**
News-Different	57.09 ± 3.59	56.62 ± 3.64	57.24 ± 3.59	58.33 ± 3.40	61.93 ± 3.02	61.36 ± 3.12	65.81 ± 2.84	**71.17 ± 2.06**
News-Same	43.47 ± 3.62	43.71 ± 3.60	43.09 ± 3.55	44.03 ± 3.43	46.21 ± 3.28	45.39 ± 3.35	48.34 ± 3.12	**56.73 ± 2.41**
News-Similar	44.40 ± 3.50	46.83 ± 3.29	44.83 ± 3.39	45.17 ± 3.41	49.54 ± 3.27	51.97 ± 2.85	50.62 ± 2.84	**56.61 ± 2.55**
USPS3568	45.05 ± 3.42	44.90 ± 3.35	44.86 ± 3.35	46.20 ± 3.26	48.59 ± 3.11	48.05 ± 3.09	**49.25 ± 2.86**	47.34 ± 3.10
USPS	63.68 ± 2.49	56.32 ± 2.88	63.90 ± 2.46	64.25 ± 2.45	67.68 ± 2.27	66.27 ± 2.29	**68.60 ± 2.10**	68.45 ± 2.08

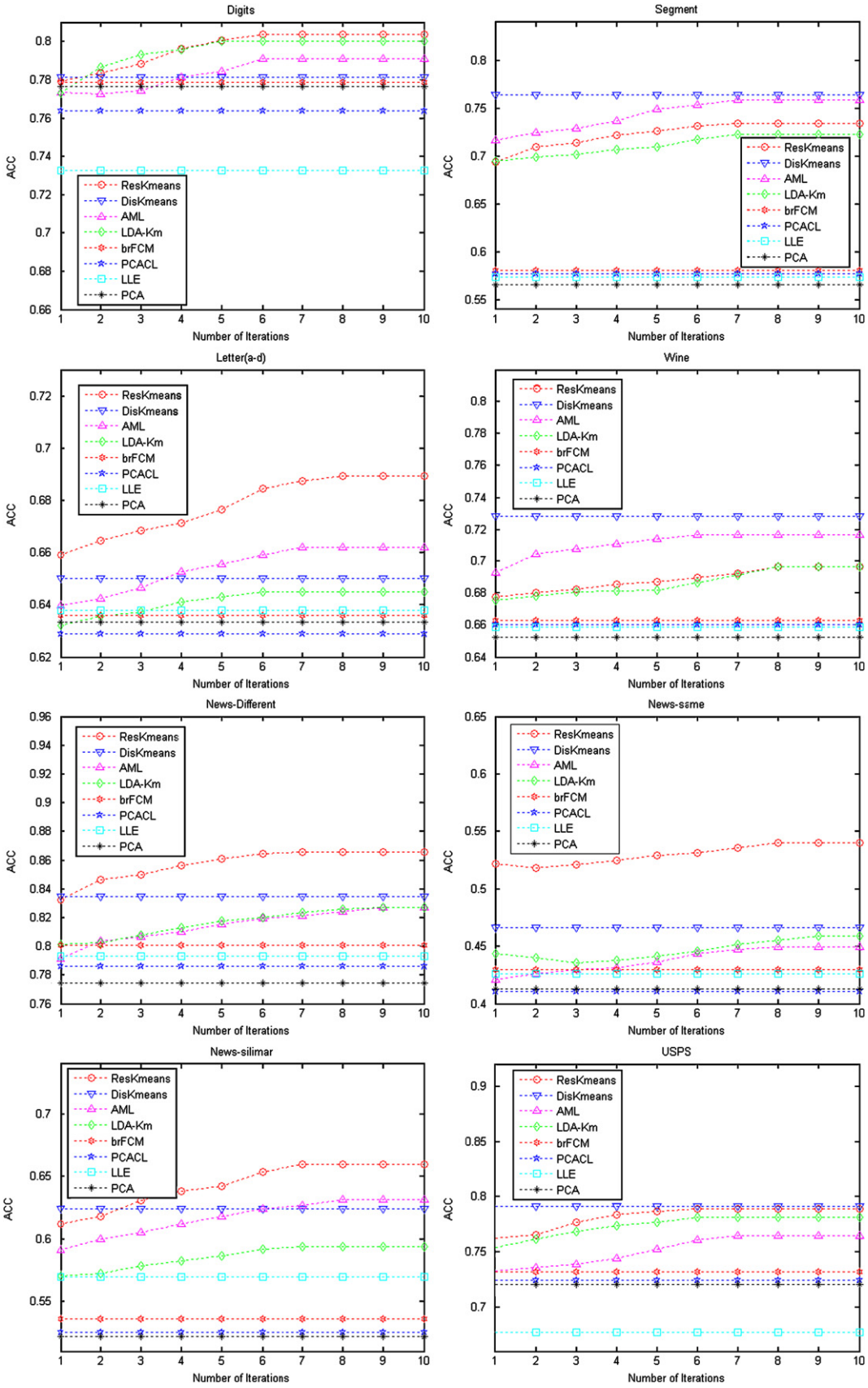

Fig. 2. ACC with various iterations.

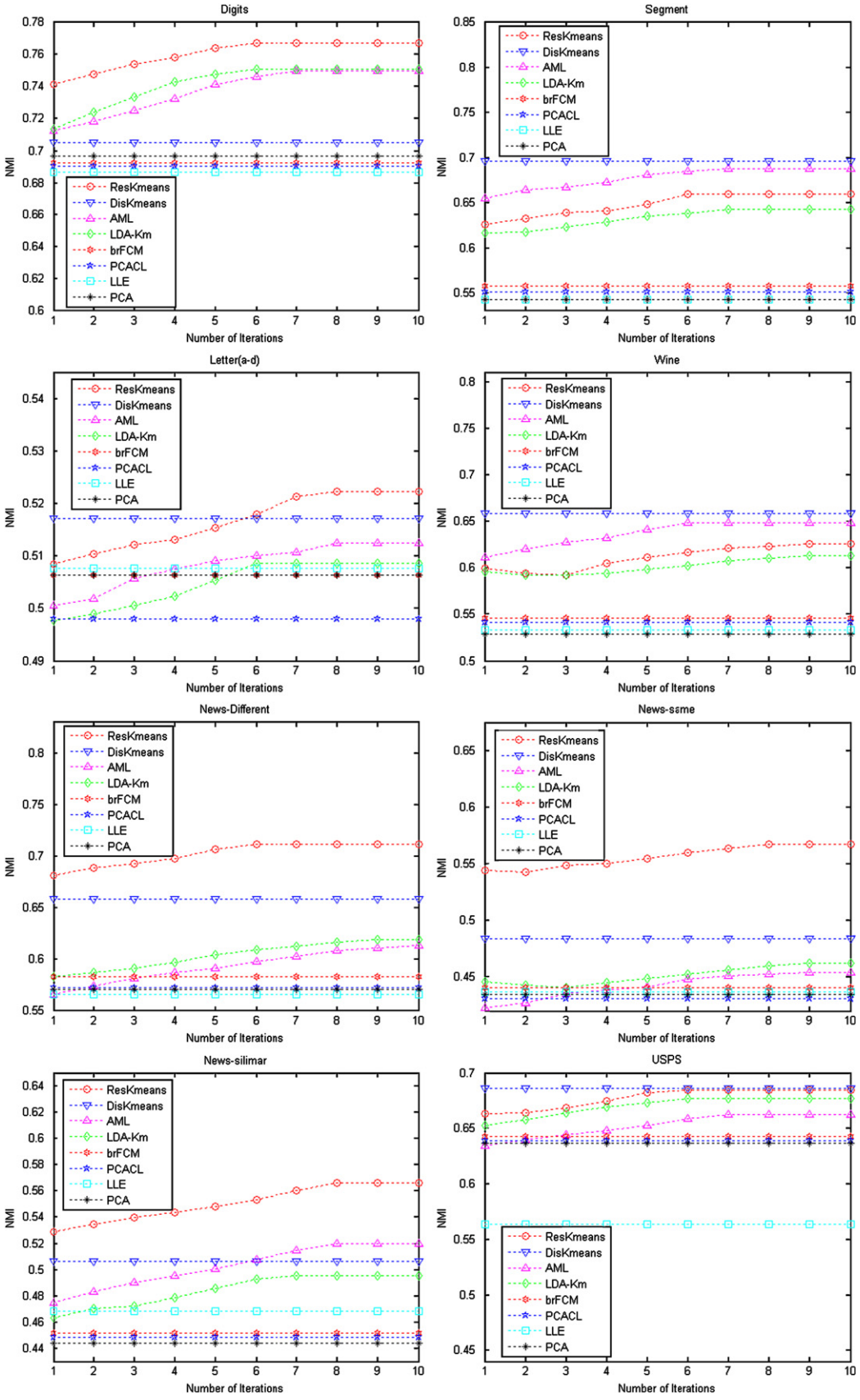

Fig. 3. NMI with various iterations.

other methods. Thus, ResKmeans is also able to perform well on the high-dimensional data sets.

From these results, we can make several interesting observations as follows:

(1) ResKmeans achieves the best performance on the ten data set used in this paper. Moreover, it has comparable performance with both DisKmeans and LDA-Km on the rest of the fifteen data sets, such as Segment, Iris, USPS and Wine. Thus, ResKmeans provides an appealing clustering performance.
(2) DisKmeans achieves relatively higher performance than LDA-Km, although they also apply the hard K-means to cluster the data in the discriminative subspace, which can be explained by the fact that DisKmeans makes effective use of the specific kernel Gram matrix optimized by the LDA criterion [15]. Additionally, in all discriminative clustering algorithms, AML performs worst on the data sets used here. The reason may be that the subspace obtained by AML is not necessarily the most discriminative one. In fact, AML only takes the separability between different clusters, and ignores the compactness within the cluster.
(3) Although DisKmeans performs considerably worse than ResKmean, it does not need to iteratively do subspace selection and clustering as shown in Figs. 2–4. On the other hand, we see that, except for DisKmeans, several discriminative algorithms can increasingly improve their clustering performances with additional iterations, and converge within about ten iterations. In contrast, the clustering performance improvement brought by ResKmeans is substantially more stable and consistent, across different data sets. This suggests that ResKmeans is adaptive to both soft clustering and

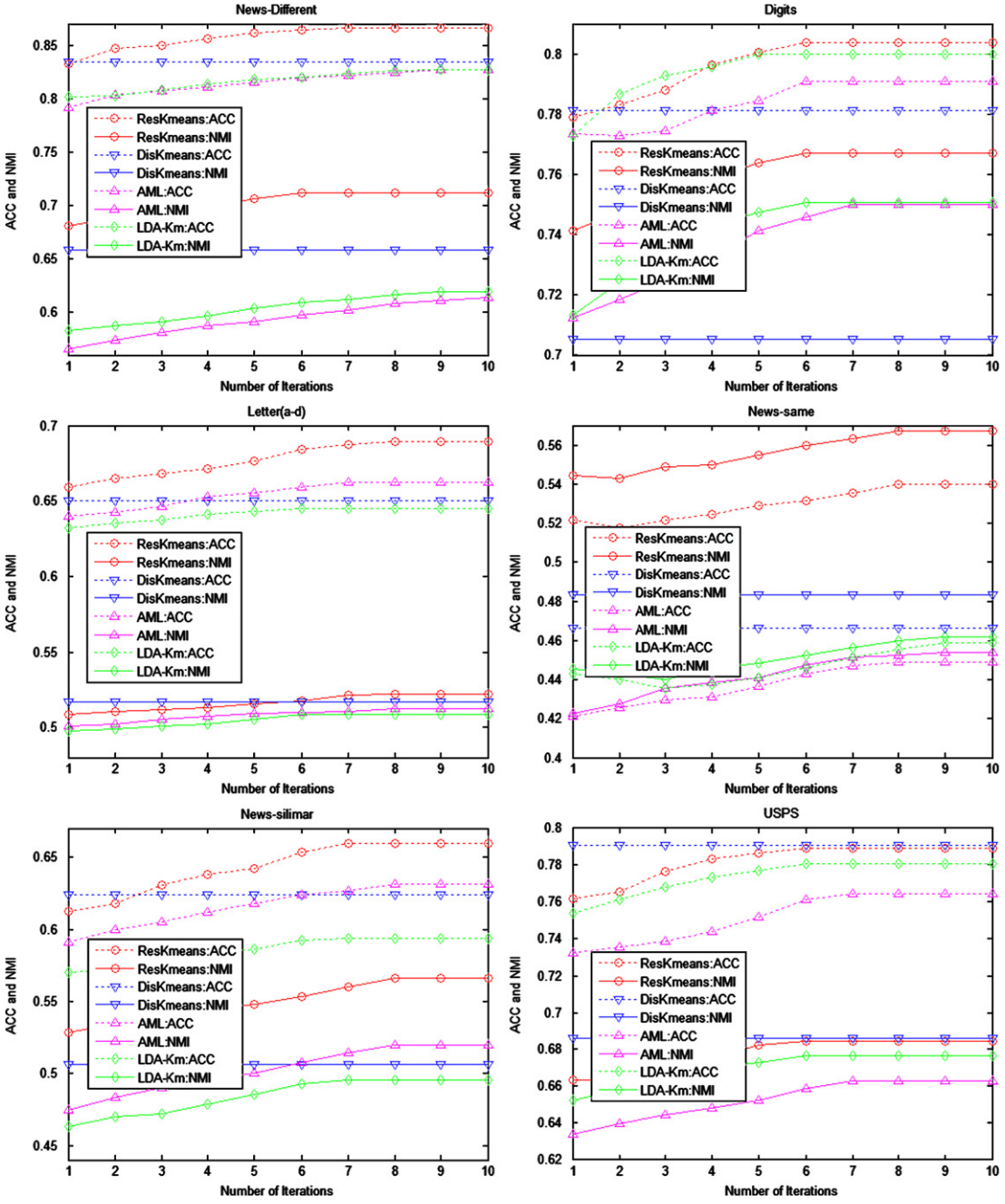

Fig. 4. Comparisons of discriminative algorithms by ACC and NMI measures.

subspace selection: it takes the feedback from the result of soft clustering to select the projection, then clusters the data in the most discriminative subspace until alternating such a process tends to convergence. Thus, we address issues (1)–(3) in the beginning of this section.

(4) As can be seen, in wherever original and corresponding reduced spaces, brFCM outperforms PCA, LLE and PCACL in most of the datasets, which indicates that the results obtained by soft clustering are better than those of hard clustering indeed. However, it is also found inferior to LDA-Km, AML, DisKmeans and ResKmeans, which indicates that the alternating methods between clustering and dimensionality reduction have relatively better performance than non-alternating ones such as brFCM performing in PCA-reduced space.

(5) We can observe from Fig. 5 that on Digits data set, the corresponding performances of all algorithms are almost unchanged with respect to the dimensions of the data from 9 to 14 ($K=10$). In other words, the eight algorithms have performed best in the 9-dimension space. Similar observation can be obtained on USPS. Secondly, on Yale, when the dimension is reduced to 14 for $K=15$, both Reskmeans and LDA-Km have obtained the highest accuracy. Although the other algorithms do not acquire their best results in the same subspace, their performances seem also not to be substantially improved as the reduced dimension is increased. For example, the cluster accuracy of DisKmeans is 0.7788 in the 14-dimension space, while 0.7824 in the 26-dimension space, their difference is not so significant and at the same time, through increasing the reduced dimension gradually, its cluster accuracy still keeps relatively constant. Analogously, we can obtain similar observation from ORL. On the whole, we can observe from Fig. 5 that although the subspace with K-1 dimensions does not guarantee optimal performance of all algorithms on a few data sets, their effective performances can still be achieved. Thus, such a performance evaluation on the above (reduced) subspace is still relatively fair for algorithmic comparison. This may be the reason why literatures [1,9,14–16,29] projected the data onto the (K-1)-th dimension subspace for clustering.

4.4. Regularization parameter

As mentioned in Section 3.1, the regularization parameter η is introduced to improve the performance of ResKmeans. In the subsection, we study its effect on the clustering performance of ResKmeans with the following experiment. In the experiment, we try a series of different η values from 10^{-5} to 10^{1} with multiplicity of 10, and obtain their corresponding results as shown in Fig. 6. We can observe that maximum entropy helps to improve the performance of ResKmeans. For example, on most of the data sets, the performance of ResKmeans with $\eta=0.01$ is significantly better than that with no regularization (i.e., $\eta=0$). Furthermore, we can see that the relatively better results can be obtained when the η value lies in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ in most cases.

5. Conclusion

In this paper, we propose a new (entropy) regularized soft K-means for discriminant analysis (ResKmeans) which combines subspace selection and clustering into a general framework.

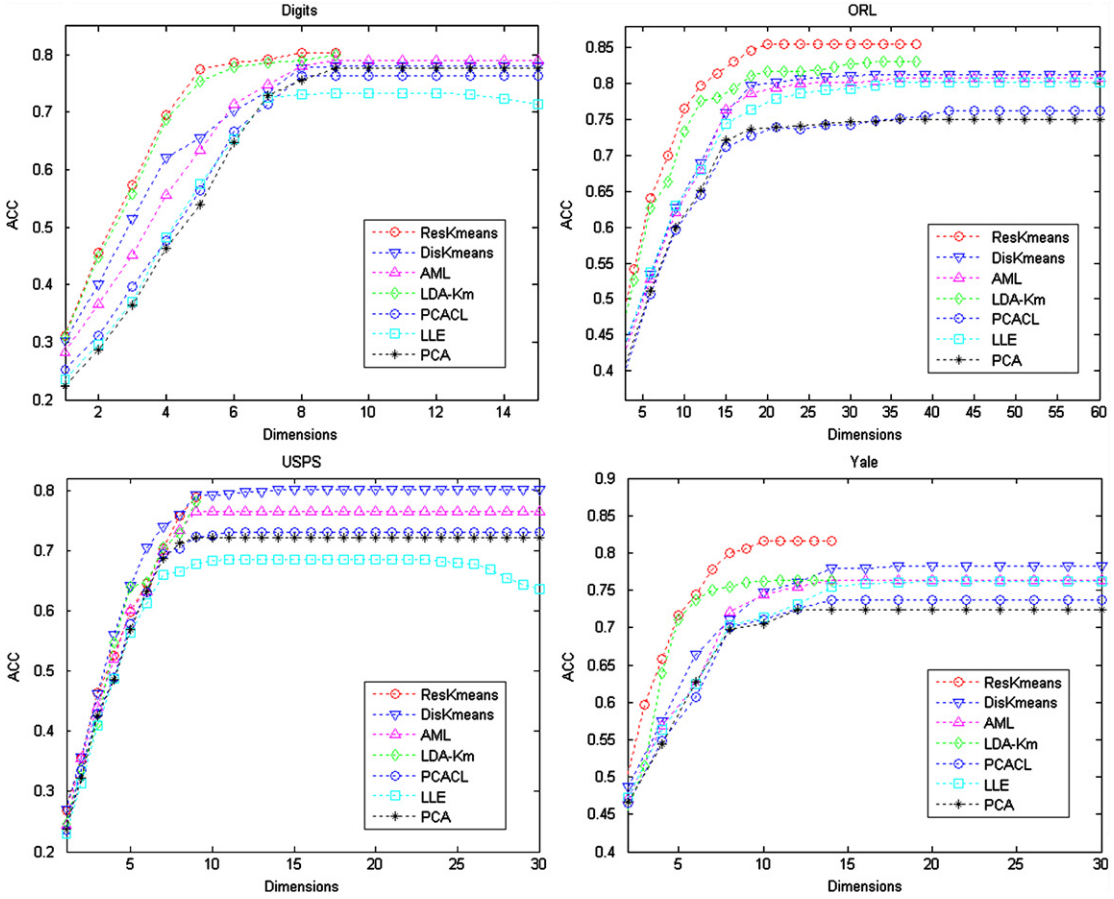

Fig. 5. Different number of dimensions.

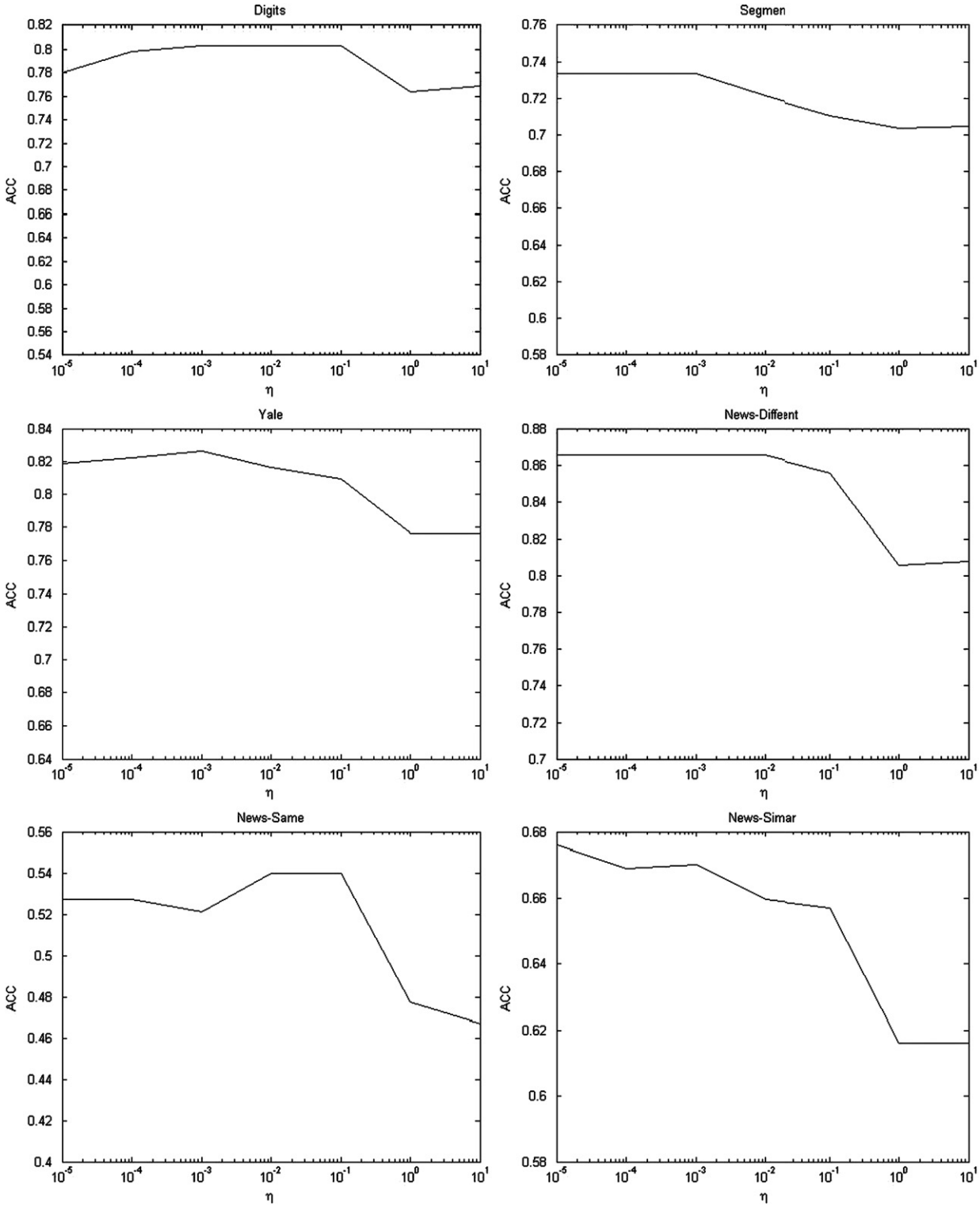

Fig. 6. The performance of ResKmeans with different η values.

Specifically, ResKmeans uses regularized soft K-means to cluster the data on the low-dimensional space to generate the membership degree of each instance; and then through making effective use of the memberships, the (soft) within-cluster scatter and (soft) between-cluster scatter matrices are obtained respectively such that we may construct generalized linear discriminant analysis (GELDA) which adaptively selects the most discriminative subspace. ResKmeans alternately performs the procedure of clustering the data points and finding the most discriminative subspace using GELDA until the appropriate cluster assignments are generated. Experimental results on real-world data sets demonstrated the effectiveness of the proposed algorithm.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to significantly improve the

quality of this paper. Also, we thank Chris Ding for useful discussions and also thank Zheng Zhao for providing the codes of AML and DisKmeans. The research is respectively supported by National Science Foundations of China and Jiangsu Province under Grant nos. 60773061 and BK2008381, Natural Science Foundation of Zhejiang Province under Grant no. Y1100349, and Doctoral Foundation of Zhejiang Radio and TV University.

Appendix

Let S_w^r, S_b^r and S_t^r respectively have the following expressions:

$$S_w^r = \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i - m_k^r)(x_i - m_k^r)^T,$$

$$S_b^r = \sum_{k=1}^{K} n_k^r (m_k^r - m^r)(m_k^r - m^r)^T,$$

$$S_t^r = \sum_{i=1}^{n} (x_i - m^r)(x_i - m^r)^T.$$

Then $S_t^r = S_w^r + S_b^r$.

Proof: We first reduce the within-cluster scatter matrix S_w^r:

$$\begin{aligned}
S_w^r &= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik}(x_i - m_k^r)(x_i - m_k^r)^T\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik}(x_i x_i^T - x_i m_k^{rT} - m_k^r x_i^T + m_k^r m_k^{rT})\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - \sum_{i=1}^{n} u_{ik} x_i m_k^{rT} - \sum_{i=1}^{n} u_{ik} m_k^r x_i^T + \sum_{i=1}^{n} u_{ik} m_k^r m_k^{rT}\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - \sum_{i=1}^{n} u_{ik} x_i m_k^{rT} - \sum_{i=1}^{n} u_{ik} m_k^r x_i^T + \sum_{i=1}^{n} u_{ik} m_k^r m_k^{rT}\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - \sum_{i=1}^{n} u_{ik} x_i m_k^{rT} - \sum_{i=1}^{n} u_{ik} m_k^r x_i^T + n_k^r m_k^r m_k^{rT}\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - n_k^r m_k^r m_k^{rT}\right) \qquad (32)
\end{aligned}$$

Also, the within-cluster scatter matrix S_b^r can be reduced as follows:

$$\begin{aligned}
S_b^r &= \sum_{k=1}^{K} n_k^r (m_k^r - m^r)(m_k^r - m^r)^T\\
&= \sum_{k=1}^{K} (n_k^r m_k^r m_k^{rT} - n_k^r m_k^r m^{rT} - n_k^r m^r m_k^{rT} + n_k^r m^r m^{rT}) \qquad (33)
\end{aligned}$$

According to Eqs. (32) and (33), we have

$$\begin{aligned}
S_b^r + S_w^r &= \sum_{k=1}^{K} (n_k^r m_k^r m_k^{rT} - n_k^r m_k^r m^{rT} - n_k^r m^r m_k^{rT} + n_k^r m^r m^{rT})\\
&\quad + \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - n_k^r m_k^r m_k^{rT}\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - n_k^r m_k^r m^{rT} - n_k^r m^r m_k^{rT} + n_k^r m^r m^{rT}\right)\\
&= \sum_{k=1}^{K}\left(\sum_{i=1}^{n} u_{ik} x_i x_i^T - \sum_{i=1}^{n} u_{ik} x_i m^{rT} - \sum_{i=1}^{n} u_{ik} m^r x_i^T + \sum_{i=1}^{n} u_{ik} m^r m^{rT}\right)\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i x_i^T - x_i m^{rT} - m^r x_i^T + m^r m^{rT})\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}(x_i x_i^T - x_i m^{rT} - m^r x_i^T + m^r m^{rT})\\
&= \sum_{k=1}^{K} u_{ik}\sum_{i=1}^{n} (x_i - m^r)(x_i - m^r)^T\\
&= \sum_{i=1}^{n} (x_i - m^r)(x_i - m^r)^T = S_t^r
\end{aligned}$$

This completes the proof.

References

[1] C. Ding T. Li. Adaptive dimension reduction using discriminant analysis and *K*-means clustering. in: Proceedings of the ICML 2007, Oregon, USA, 2007, pp. 521–528.
[2] J. Tenenbaum, V.D. Silva, J. Langford., A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[3] C. Chen, L. Zhang, J. Bu, C. Wang, Wei Chen, Constrained Laplacian Eigenmap for dimensionality reduction, Neurocomputing 73 (2010) 951–958.
[4] I. Jolliffe, Principal Component Analysis, 2nd edition, Springer, 2002.
[5] L. Ezequiel, O.J. Miguel, M. José, A.G. José, Principal components analysis competitive learning, Neural Comput 16 (2004) 2459–2481.
[6] L.K. Saul, S.T. Roweis, Think globally, fit locally: Unsupervised learning of low dimensional manifolds, J. Mach. Learn Res 4 (2003) 119–155.
[7] X.F. He, P. Niyogi, Locality Preserving Projections. in: Proceedings of the Advances in Neural Information Processing Systems (NIPS 2003), Vancouver, Canada, 2003.
[8] B. Yang, S. Chen, Sample-dependent graph construction with application to dimensionality reduction, Neurocomputing 74 (2010) 301–314.
[9] J.P. Ye, Z. Zhao, H. Liu. Adaptive distance metric learning for clustering. In: Proceedings of the CVPR 2007, Minneapolis, USA, 2007, pp. 1–7.
[10] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell 19 (1997) 711–720.
[11] L. Yang, R. Jin, L.B. Mummert, R. Sukthankar, A. Goode, B. Zheng, C.H. Hoi, M. Satyanarayanan., A. Boosting, Framework for visuality-preserving distance metric learning and its application to medical image retrieval, IEEE Trans. Pattern Anal. Mach. Intell 32 (2010) 30–44.
[12] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin., Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell 29 (2007) 40–51.
[13] L. Sun, S. Ji, J. Ye, Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis., IEEE Trans. Pattern Anal. Mach. Intell 33 (2011) 194–200.
[14] F. De la Torre, T. Kanade. Discriminative cluster analysis. In: Proceedings of the ICML 2006, Pittsburgh,USA, 2006, pp. 241–248.
[15] J.P. Ye, Z. Zhao, M.R. Wu, Discriminative *K*-means for Clustering. In: Proceedings of the Advances in Neural Information Processing Systems (NIPS 2007), Vancouver, B.C., Canada, 2007, pp. 1649–1656.
[16] J.H. Chen, Z. Zhao, J.P. Ye, H. Liu. Nonlinear adaptive distance metric learning for clustering. In: Proceedings of the SIGKDD 2007, California, USA, 2007, pp. 123–132.
[17] S. Miyamoto, M. Mukaidono. Fuzzy c-means as a regularization and maximum entropy approach. In: Proceedings of the IFSA 1997, Prague, Czech Republic, 1997, pp. 86–92.
[18] W.L. Cai, S.C. Chen, D.Q. Zhang., Fast and robust Fuzzy C-means clustering algorithms incorporating local information for image segmentation, Patt. Recog 40 (2007) 825–838.
[19] S. Eschrich, J. Ke, L.O. Hall, D.B. Goldgof., Fast accurate fuzzy clustering through data reduction, IEEE Trans. Fuzzy Syst 11 (2003) 262–270.
[20] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell 11 (1989) 773–778.
[21] D.Q. Zhang, S.C. Chen, Kernel-based fuzzy and possibilistic c-means clustering. In: Proceedings of the ICANN 2003, Istanbul, Turkey, 2003, pp. 122–125.
[22] W.W. Du, K.H. Inoue, K. Urahama, Robust kernel fuzzy clustering, Fuzzy Syst. Knowl. Discovery 36 (2005) 454–461.
[23] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[24] S. Eschrich, J. Ke, L.O. Hall, D.B. Goldgof., Fast Accurate Fuzzy Clustering Through Data Reduction, IEE, Trans. Fuzzy Syst 11 (2003) 262–270.
[25] R.N. Dave, S. Sen, Robust fuzzy clustering of relational data, IEEE Trans. Fuzzy Syst 10 (2002) 713–727.
[26] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc 84 (1989) 165–175.
[27] X. Yin, S. Chen, E. Hu, D. Zhang., Semi-supervised clustering with metric learning: An adaptive kernel method, Pattern Recognition 43 (2010) 1320–1333.
[28] D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering, IEEE Trans. Knowl. Data Eng 23 (2011) 902–913.
[29] W. Tang, H. Xiong, S. Zhong, J. Wu. Enhancing semi-supervised clustering: a feature projection perspective. In: Proceedings of the KDD 2007, San Jose, California,USA, 2007, pp. 707–716.

Xuesong Yin received the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010.

He is currently an Associate Professor in the School of Information and Engineering, Zhejiang Radio and TV University, Hangzhou, China. His current research interests include machine learning, data mining, and pattern recognition.

Songchan Chen received the B.Sc. degree in mathematics from Hangzhou University, Hangzhou, China, in 1983, the M.Sc. degree in computer applications from Shanghai Jiaotong University, Shanghai, China, in 1985, and the Ph.D. degree in communication and information systems from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1997.

He was an Assistant Lecturer at NUAA in January 1986. Since 1998, he has been with the Department of Computer Science and Engineering at NUAA as a full Professor. He has authored or co-authored over 130 scientific papers. His current research interests include pattern recognition, machine learning, and neural computing.

Enliang Hu received the M.Sc. degree in mathematics from Yunnan Normal University, Kunming, China, in 2003, and the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010.

He is currently an Associate Professor in the Department of Mathematics, Yunnan Normal University. His current research interests include machine learning, pattern recognition, and data mining.