Pattern Recognition Letters xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A subspace co-training framework for multi-view clustering ☆

Xuran Zhao *, Nicholas Evans, Jean-Luc Dugelay

EURECOM, Multimedia Communication Department, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Multi-view clustering
Subspace clustering
Co-training

ABSTRACT

This paper addresses the problem of unsupervised clustering with multi-view data of high dimensionality. We propose a new algorithm which learns discriminative subspaces in an unsupervised fashion based upon the assumption that a reliable clustering should assign same-class samples to the same cluster in each view. The framework combines the simplicity of k-means clustering and Linear Discriminant Analysis (LDA) within a co-training scheme which exploits labels learned automatically in one view to learn discriminative subspaces in another. The effectiveness of the proposed algorithm is demonstrated empirically under scenarios where the conditional independence assumption is either fully satisfied (audio-visual speaker clustering) or only partially satisfied (handwritten digit clustering and document clustering). Significant improvements over alternative multi-view clustering approaches are reported in both cases. The new algorithm is flexible and can be readily adapted to use different distance measures, semi-supervised learning, and non-linear problems.

1. Introduction

The recent explosion of multimedia information on the Internet demands effective clustering techniques capable of handling huge quantities of potentially complex data. First, multimedia data are generally represented in high-dimensional spaces in which the so-called *curse-of-dimensionality* makes the application of many clustering techniques somewhat troublesome. Second, by its very nature, multimedia data is multi-modal, for example audio and video information can form two independent clustering inputs. The fusion of modalities remains a challenging problem and is generally treated in isolation to that of high dimensionality.

Difficulties associated with the high dimensionality are generally overcome through the application of dimensionality reduction (DR) techniques, such as Principle Component Analysis (PCA) (Jolliffe, 2005) and related approaches. Dimensionality reduction can either be applied in a pre-processing step prior to clustering, or be integrated into the clustering framework itself. The latter is referred to as subspace clustering (see a survey (Kriegel et al., 2009)). Whatever the technique, however, the goal is always to identify a subspace in which clusters are maximally separated.

Research in multi-modal fusion, which aims to optimally combine information in different views of the same data, has led to a number of multi-view clustering algorithms, e.g. (Bickel and Scheffer, 2004; Chaudhuri et al., 2009; Kumar and Daumé, 2011). The goal with all such methods is to identify a clustering result which agrees across different views (samples clustered together in one view are also clustered together in other views).

This paper presents our efforts to address the problems of high-dimensionality and multi-modal fusion in a unified framework. We assume that each data sample is represented by two feature vectors corresponding to two independent views. We further assume significant information in each feature vector to be unrelated to the underlying class label and that there exists a lower dimensional subspace in which classes are maximally separated. Inspired by the concept of *co-training* (Blum and Mitchell, 1998), we describe a new multi-view subspace clustering algorithm which reflects the intuition that a true underlying clustering should assign samples to the same cluster irrespective of the view. It seeks a discriminant subspace for each view which results in a clustering policy with maximal agreement across views. Discriminant subspaces in one view are learned using cluster labels for the same samples in another view, and vice versa. The process is iterative and is repeated until a maximum agreement is achieved. The proposed algorithm simultaneously outputs cluster indicators, discriminant subspaces for each view, and compact models of different clusters. As a result, the algorithm copes naturally with out-of-sample data and is readily extended to semi-supervised classification.

The remainder of this paper is organized as follows. Section 2 analyses three essential components of the proposed algorithm: LDA, k-means, and co-training. Section 3 presents the proposed clustering algorithm and extensions to cosine distance, non-linear case and semi-supervised settings. Section 4 describes the proposed algorithm in the context of existing literature. Section 5

☆ This paper has been recommended for acceptance by Jesús Ariel Carrasco Ochoa.

* Corresponding author. Tel.: +358 41 4996553.
E-mail addresses: zhaox@eurecom.fr, xuran.zhao@eurecom.fr (X. Zhao), evans@eurecom.fr (N. Evans), dugelay@eurecom.fr (J.-L. Dugelay).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.12.003

presents experimental evaluations in audio-visual speaker clustering. Section 6 presents our conclusions.

2. LDA, k-means, and co-training

In this section we describe the three essential components of the proposed algorithm: LDA, k-means and co-training.

2.1. LDA and k-means

As discussed in Ding and Li (2007), the objective function of LDA and k-means are closely related. Consider a set of centered input data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ such that $\bar{\mathbf{x}} = \sum_i \mathbf{x_i}/n = 0$. Let the class labels be given by $H = \{h_1, \ldots, h_n\}$, and define matrices of between-class scatter S_b, within-class scatter S_w and total scatter S_t as:

$$\begin{aligned} S_b &= \sum_k n_k \mathbf{m}_k \mathbf{m}_k^T \\ S_w &= \sum_k \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \\ S_t &= \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \end{aligned} \tag{1}$$

where n_k is the number of samples in class k, $\mathbf{m}_k$ is the mean of class k, and C_k is the set of samples belonging to kth class ($l_i = k$) and $S_t = S_w + S_b$. LDA seeks a projection P which maximizes the ratio between S_b and S_w. The objective function is thus:

$$\begin{aligned} \arg\max_P \operatorname{tr} \frac{p^t S_b p}{p^t S_w p} &= \arg\max_P \operatorname{Tr} \frac{P^T S_b P}{P^T S_w P} + 1 = \arg\max_P \operatorname{Tr} \frac{P^T S_t P}{P^T S_w P} \\ &= \arg\min_P \operatorname{Tr} \frac{P^T S_w P}{P^T S_t P} \end{aligned} \tag{2}$$

Where $\operatorname{Tr}\{\cdot\}$ is the trace of a matrix.

On the other hand, the k-means objective function is give by:

$$\arg\min_H \sum_k \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2 \tag{3}$$

where H represents a cluster indicator and $\mathbf{m_k}$ is the mean of the kth cluster. In most cases same-class samples should be assigned to the same cluster, i.e. cluster labels should be indicative of the class label L. In this case, the k-means objective function is equivalent to the minimization of the trace of the within-class scatter matrix so that:

$$\arg\min_H \operatorname{Tr} S_w = \arg\min_H \operatorname{Tr} (S_t - S_b) \tag{4}$$

Eqs. (2) and (4) thus reveal that the LDA and k-means objective functions are compatible: k-means aims to minimize within-class scatter while LDA minimizes the within-class scatter and maximize total scatter in the same time.

2.2. Co-training

Co-training (Blum and Mitchell, 1998) is one of the most acclaimed approaches to semi-supervised learning. In co-training, data samples are assumed to be represented by two conditionally independent features X_1 and X_2. Two predictors f_1 and f_2 assign to each X a class label Y ($f : X \to Y$) and are trained according to each view using a small pool of labeled data. The two predictors are used to assign labels to a larger pool of unlabeled data. A subset of samples with which the predictors have the most confidence in label assignments is added to the pool of labeled data. The predictors are then iteratively re-learned and applied to the remaining unlabeled data. Co-training essentially learns two different predictors f_1 and f_2 which *agree* on unlabeled data across different views. A theoretical treatment of convergence is given in the original paper Blum and Mitchell (1998) and shows that, under the assumption of conditional independence, a weak predictor f_1 in view X_1 which can tolerate random label noise can learn from automatically labeled samples provided by f_2 in view X_2.

This paper presents the extension of co-training predictors to co-training subspaces. LDA is a supervised method which requires class labels, while k-means is a unsupervised method which generates cluster indicators. Under the assumption of conditional independence between views, they can be regarded as class labels corrupted with random noise for the other view. The two methods are combined with the idea of co-training.

3. Multi-view subspace clustering: a co-training algorithm

In this section, we apply the concept of co-training to the problem of discriminant subspace learning for multi-view clustering. Since we assume unsupervised clustering, the standard semi-supervised co-training algorithm cannot be applied directly. However, the goal remains the same, i.e. to learn a subspace for each view which results in a common clustering policy. For clarity, samples assigned to the same cluster in the subspace of one view should be assigned to the same cluster in the subspace of the other view and, conversely, samples assigned to different clusters in the subspace of one view should be assigned to different clusters in the subspace of the other view.

3.1. An algorithm: CoKmLDA

We first define a *Cluster Agreement Index* (CAI). Let $H^{(1)}$ and $H^{(2)}$ represent the assignment of samples in views $v = 1$ and $v = 2$ to one of K clusters. The CAI is defined as:

$$CAI(H^{(1)}, H^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} \delta\left(h_i^{(1)}, \operatorname{map}(h_i^{(2)})\right) \tag{5}$$

where n is the total number of samples and $\delta(a, b)$ is a function equal to unity if $a = b$ and zero otherwise. The $map(\,)$ function returns an optimal mapping between cluster identifiers in view 1 to those in view 2 in order that the CAI is maximized. This is achieved with a classical Hungarian algorithm (Steiglitz and Papadimitriou, 1982).

We then seek two LDA projections $P^{(1)}$ and $P^{(2)}$ such that the CAI resulting from k-means on both subspaces is maximized. The objective function is given by:

$$\arg\max_{P^{(1)}, P^{(2)}} CAI(H^{(1)}, H^{(2)}) \tag{6}$$

where $H^{(v)}$s are further dependent on $P^{(v)}$s

$$H^{(v)} = \arg\min_{H^{(v)}} \sum_{k=1}^{K} \sum_{h_i^{(v)} = k} \|P^{(v)^T} \mathbf{x_i} - P^{(v)^T} \mathbf{m_k}\|^2 \ (v = 1, 2) \tag{7}$$

In the following we propose an algorithm that alternatively solves Eqs. (6) and (7) for $P^{(v)}$ and $H^{(v)}$ according to a modified co-training approach. We use cluster indicators generated by k-means in one view as label information to train LDA projections in the other view, and vis-versa. While the essential elements of the proposed algorithm are relatively straightforward, the algorithm tends to converges given that LDA can learn approximately good projections with some extent of label noise (mathematical proof given in Section 3.3). The new algorithm is referred to as co-k-means Linear Discriminant Analysis (CoKmLDA). The main steps of the iterative algorithm are as follows:

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

Algorithm 1. CoKmLDA

Input: a set of n multi-view samples $X = \{X^{(v)} | v = 1, 2\}$, where $X^{(v)} = \{x_1^{(1)}, \ldots, x_n^{(v)}\}$, and the expected number of clusters K.

Output: view dependent cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$, and projection matrices $P^{(1)}, P^{(2)}$

Initialize:

1. Center the feature vectors in each view and apply PCA if the dimensionality of the feature space is too high;
2. Perform k-means clustering in each view to estimate cluster indicators $H^{(v)} = \{h_1^{(v)}, \ldots, h_n^{(v)}\}$;
3. For each view v, identify the single sample closest to each of the K clusters, $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.

for $t = 1$ **to** *iter* **do**

for $v = 1$ **to** 2 **do**

1. Use $X^{(v)}$ and $H^{(3-v)}$ to train LDA projections $P^{(v)}$ and project samples into the LDA subspace;
2. Using seeds $S^{(v)}$, perform k-means clustering on projected samples to estimate new cluster indicators $H^{(v)}$;
3. Update seeds $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$.

end for

end for

1. **k-means clustering** Solve Eq. (7) with fixed $P^{(1)}$ and $P^{(2)}$ by determine cluster indicators $H^{(1)}$ and $H^{(2)}$ with a k-means algorithm operating in each view. In the initialization step, k-means is applied on original features. If the dimensionality of the original feature is too high, PCA is applied as a preprocessing step.
2. **Cross-labeling** Label samples in view 1 according to $H^{(2)}$, and vis-versa.
3. **LDA training** Learn LDA projection $P^{(1)}$ with original or PCA processed features $X^{(1)}$ and labels corresponding to view 2, and vis-versa. This step optimizes Eq. (6) in the sense that, in view 2, samples belongs to the same cluster indicated by $H^{(1)}$ will be projected near each other while samples belongs to different clusters indicated by $H^{(1)}$ will be projected apart. So the data structure in view 2 will be constrained to be more compatible with view 1, vis-versa.
4. **Iterate** Return to step 1, perform k-means again in projected subspace. We compute the objective function in Eq. (6) for each iteration. The iteration process can be terminated either after a fixed number of iterations, or when the objective function did not reach a new minimum for a fixed number of iterations in each view.

It is well known that the performance of k-means is sensitive to the quality of its initialization (seeds). Accordingly it is common to run k-means several times with random initialization, and to retain the clustering result which minimizes Eq. (4). This approach is computationally demanding and thus we utilize *seed inheritance* to reduce the computational burden. After each application of k-means, we identify in each view the single sample closest to each of the K cluster centroids, denoted $S^{(v)} = \{s_1^{(v)}, \ldots, s_K^{(v)}\}$. In subsequent iterations, k-means applied in view v is initialized with the K seeds in $S^{(v)}$. The CoKmLDA algorithm is formally summarized in Algorithm 1.

The computational complexity of single iteration is in the order of $O(pn)$ for k-means, and $O(p^2 n)$ for LDA, where p is the feature dimensionality and n is the number of samples. For t iterations the complexity of CoKmLDA algorithm is hence $O(pnt + p^2 nt)$.

3.2. An illustrative example

Here we illustrate the behavior and merits of the proposed CoKmLDA algorithm using synthetic data of 300 samples represented in two views, each of two dimensions. Each sample belongs to one of two classes, where each class is a two component Gaussian mixture. All four Gaussian components have a covariance matrix of $\Sigma = diag(0.3, 0.3)$. The means of each Gaussian component are detailed in Table 1. The two views are conditionally independent, i.e. two samples generated by the same Gaussian component in one view can belong to different Gaussian components in the other view. Finally, the number of samples corresponding to each Gaussian components, also illustrated in Table 1, is intentionally unbalanced in order that, for initialisation, k-means gives better-than-random accuracy relative to real class labels.

Scatter plots of generated data in 2 views are show in Fig. 1(a). Fig. 1(b) illustrates clustering results after the initial application of k-means in the original feature space. We note a high degree of error; two of the four Gaussian components are incorrectly clustered. The result of cross-labeling, where samples in each view are labeled according to the clustering indicators in the other view, is shown in Fig. 1(c). The two crosses at center of each plot represents the two cluster centroids in each view. The resulting LDA projections (1-dimensional for this trivial example) are shown by the solid, straight lines in Fig. 1(c). After the samples are projected into the new subspaces and k-means is reapplied, the clustering results are greatly improved as illustrated in Fig. 1(d). The new cluster centroids and LDA projections normally used in the second iteration are also illustrated in Fig. 1(d). Whereas several iterations are required in practice, the new clustering result is fully representative of the true underlying class structure and the algorithm converges in a single iteration for this illustrative example.

3.3. Mathematical analysis

The above example illustrates the behavior of the algorithm for a trivial example. Given the assumption of conditional independence between the two views, clustering indicators in one view can be utilized as class labels in another view, but with random label noise. Here we aim to show mathematically that *LDA projection can be learned with labeled samples with random label noise*.

We first define a hypothetical level of label noise λ. Let there be n centered data samples, $X = [x_1, \ldots, x_n]$ and let X_k and n_k be the subset and number of samples in the kth class respectively. For each class, $(1-\lambda)n_k$ and λn_k points are randomly sampled from X_k and $X - X_k$ respectively to form a new subset X_k^* for the kth class with random label noise. In the following we show that the expected LDA projection trained with X_k^* is equivalent to the LDA projection trained with true labels.

Trained on X_k^* with noisy labels, the LDA projection P is determined according to:

$$\max_P \mathrm{Tr} \frac{P^T S_b^* P}{P^T S_t^* P} \tag{8}$$

where S_b^* and S_t^* are the between-class and total scatter estimated with noisy data. It is clear that $S_t^* = S_t$ since its calculation do not need label information, whereas S_b is defined as:

Table 1
(x, y) Locations of Gaussian centroids and number of samples.

	Class 1 (red)	Class 2 (blue)
View 1	(−2,4) 50 smpl.	(2,4) 100 smpl.
	(−4,−4) 100 smpl.	(2,−4) 50 smpl.
View 2	(−4,−2) 100 smpl.	(−4,2) 50 smpl.
	(4,−2) 50 smpl.	(4,2) 100 smpl.

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

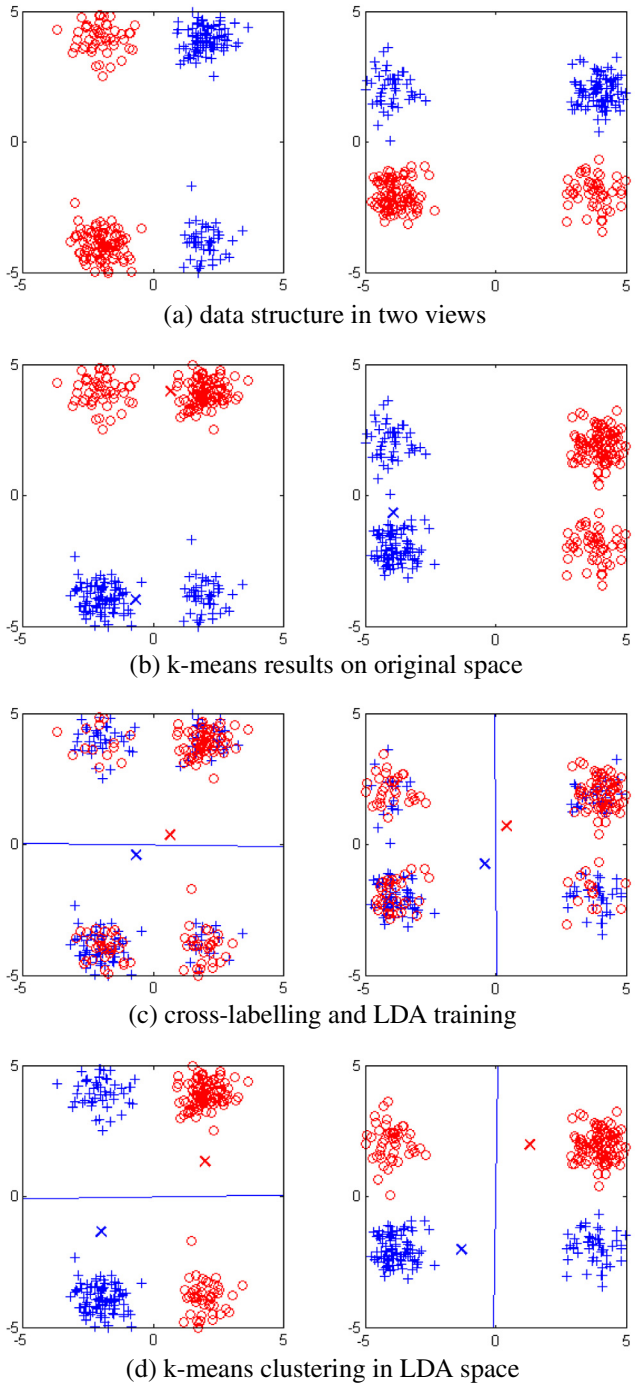


Fig. 1. Illustration of a test run of CoKmLDA on synthetic dataset.

$$S_b^* = \sum_k n_k \mathbf{m}_\mathbf{k}^* \mathbf{m}_\mathbf{k}^{*T} \tag{9}$$

where $\mathbf{m}_\mathbf{k}^*$ is the mean of X_k^*. Its value in the sense of statistical expectation is given by:

$$\mathbb{E}(\mathbf{m}_\mathbf{k}^*) = \mathbb{E}\left(\frac{1}{n_k}\left(\sum_{i=1}^{(1-\lambda)n_k} x_{ki}^+ + \sum_{i=1}^{\lambda n_k} x_{ki}^-\right)\right) = (1-\lambda)\mathbb{E}(x_{ki}^+) + \lambda\mathbb{E}(x_{ki}^-) \tag{10}$$

where x_{ki}^+ is the ith sample from X_k and x_{ki}^- is the ith sample from $X - X_k$. It is straightforward that

$$\begin{aligned} \mathbb{E}(x_{ki}^+) &= \text{mean}(X_k) = \mathbf{m}_\mathbf{k} \\ \mathbb{E}(x_{ki}^-) &= \text{mean}(X - X_k) = -\frac{n_k}{n - n_k}\mathbf{m}_\mathbf{k} \end{aligned} \tag{11}$$

Combining Eqs. (10) and (11), we obtain:

$$\mathbb{E}(\mathbf{m}_\mathbf{k}^*) = \left(1 - \frac{\lambda n}{n - n_k}\right)\mathbf{m}_\mathbf{k} \tag{12}$$

From Eq. (12) we observe that the expected value of m_k^* estimated with X_k^* containing noisy labels lies in the same direction relative to the origin as in the case where it is estimated with clean labels, but with a shorter vector norm. This can be observed in Fig. 1(c) and (d) in which the two class means in each view lie in the same direction, but different distances from the origin.

Upon substitution of Eq. (11) into Eq. (9), we obtain the expectation of S_b^*:

$$\mathbb{E}(S_b^*) = \sum\nolimits_k n_k \mathbb{E}(\mathbf{m}_\mathbf{k}^*)\mathbb{E}(\mathbf{m}_\mathbf{k}^{*T}) = \sum\nolimits_k n_k \left(1 - \frac{\lambda n}{n - n_k}\right)^2 \mathbf{m}_\mathbf{k}\mathbf{m}_\mathbf{k}^T \tag{13}$$

If we assume an equal number of sample per class, i.e. a constant $n_k = n/K$, then:

$$\mathbb{E}(S_b^*) = \left(1 - \frac{\lambda K}{K-1}\right)^2 S_b \tag{14}$$

and if S_b^* and S_t^* in Eq. (8) are replaced with their expected values, we obtain:

$$\max_P \text{Tr}\ \frac{C^2 P^T S_b P}{P^T S_t P} \tag{15}$$

where $C = (1 - \frac{\lambda K}{K-1})$ is a constant. Eq. (15) shows that LDA objective function in the case of sample with random label noise is equivalent to the objective function in the case of clean labels.

3.4. Extensions of CoKmLDA

This paper presents the idea of unsupervised subspace clustering using co-training. The framework is entirely flexible and may combine different clustering methods and supervised dimensionality reduction algorithms according to specific application and nature of related data. For example, to cluster text data, cosine distance is a more appropriate distance measure, and for non-linear separable data, kernel methods are often applied. In this section, we first presents three extensions related to clustering, namely cosine k-means, kernel approach and semi-supervised extension. We also provide multi-view extension to adapt to the situation where the data is represented by more than two views.

Cosine distance extension: The standard k-means algorithm uses a Euclidean distance metric. In some experiments in multimodal face and speaker recognition, however, we observe that the cosine distance normally gives better performance when used in LDA subspace. Tang et al. (2012) report similar findings in the context of speaker clustering. The use of a cosine distance metric in clustering problems is proposed in Dhillon et al. (2001) which reports a spherical k-means algorithm which maximizes the sum of the cosine similarity between samples and related cluster centroids. Spherical k-means follows a similar iterative process as standard k-means, except that feature vectors are first normalized to have unit length and, in the assignment step, samples are assigned to the cluster centroid which has the highest cosine similarity. The power of spherical k-means clustering is brought to CoKmLDA simply by replacing the standard k-means step in Algorithm 1.

Kernel extension: LDA learns a subspace in which classes are better separated. In the event that classes are not linearly separable in the original space, then performance is usually poor. Using a kernel trick similar to that employed in Support Vector Machines (SVM), LDA can be implicitly performed in a new feature space, which allows non-linear mappings to promote maximum separability of different classes. This approach is commonly referred to

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

as Generalized Discriminant Analysis (GDA) (Baudat and Anouar, 2000). By replacing standard LDA by GDA, the proposed algorithm may be applied to clustering problems in which multi-view data is not linearly separable.

Semi-supervised extension: The algorithm is also readily extended to semi-supervised clustering when a subset of manually labeled data in addition to a larger subset of unlabeled data are available. In this case the initial k-means step uses centroid statistics acquired from the manually labeled data as proposed in Basu et al. (2002). In our approach the k-means algorithm is seeded in each iteration with labeled data. In the case where the number of classes is high, and where random initialization often generates several seeds corresponding to some classes whereas none for others, this seemingly naive method often brings significant improvements in performance in our framework. The proposed algorithm simultaneously determines discriminant subspaces in addition to compact cluster/class models and is naturally equipped to handle out-of-sample data. Unseen test data can be projected into the relevant subspaces and classified according to the nearest centroid.

Multi-view extension: Finally, it is possible to extend the proposed two-view CoKmLDA algorithm to multi-view clustering. Assuming that each data sample is represented by m-views ($m > 2$), subsequent to the initialization and each iteration in Algorithm 1, m sets of cluster indicators are generated. In the two-view setting, an LDA projection in one view is learnt using cluster indicators in the other view to enforce a similar data structure in each subspace. Extending to an m-view setting, a straight forward solution involves the learning of an LDA projection in one view using cluster indicators of *all other views* as class labels.

Traditional LDA accepts only a single label vectors. In order to deal with multiple label vectors, the traditional LDA algorithm is modified as follows. Assume a set of centered input data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and m sets of class indicators $\{H^{(1)}, \ldots, H^{(m)}\}$. We first calculate the within-class scatter $S_w^{(v)}$ and the between-class scatter $S_b^{(v)}$ using each class indicator $H^{(v)}$ according to Eq. (1). The overall between-class scatter S_b and within-class scatter S_w are then defined as:

$$S_b = \sum_{v=1}^{m} S_b^{(v)}; \quad S_w = \sum_{v=1}^{m} S_w^{(v)}. \tag{16}$$

Finally, the optimal projection P is obtained in the same way as for the traditional LDA by optimizing the objective function in Eq. (2). Despite the different formulation, this method is have similar effect to the Multi-label Linear Discriminant Analysis (MLDA) proposed in Wang et al. (2010). The proposed method is still referred to as MLDA for simplicity.

To conclude, multi-view CoKmLDA differs from the two-view CoKmLDA in that in step 1 of the iterative process of Algorithm 1, an MLDA projection $P^{(v)}$ is learnt using $X^{(v)}$ and the cluster indicators $\{H^{(1)}, \ldots, H^{(v-1)}, H^{(v+1)}, \ldots, H^{(m)}\}$ from all other views, while all other operations remain the same.

4. Related works and analysis

Several different approaches to subspace and multi-view clustering have been reported in the open literature. Here we discuss their relationship with the new algorithm proposed in this paper.

Subspace clustering (Kriegel et al., 2009) refers to a general class of clustering methods which aim to discover a subspace more amenable to clustering. These methods are largely uni-modal. Among numerous other examples, the most relevant to the proposed algorithm are the LDA-km algorithm (Ding and Li, 2007) and DisKmeans (Ye et al., 2007) which use cluster indicators generated by k-means to learn an LDA projection. As a form of self-training, such approaches do not generally lead to significant improvements in clustering performance over a baseline k-means. The proposed CoKmLDA algorithm can be regarded as a co-training extension of Ding and Li (2007).

In the multi-view clustering setting, the general objective is to find certain kind of agreement between different views. Recent approach to multi-view clustering can be roughly divided into two major categories. The first category of algorithms is multi-view spectral clustering based on similarity graphs. As shown in Fig. 2(a), a similarity graphs (matrix) $S^{(v)}$ is first constructed for each view $X^{(v)}$ where $S_{ij}^{(v)} = exp(-\|x_i, x_j\|^2/t^2)$, where $\|\cdot\|$ is a certain distance measure and t is the Gaussian bandwidth, thus $S_{ij}^{(v)}$ represents the similarity between ith and jth sample in the vth view. The original similarity graphs $S^{(v)}$ are then transformed so that the difference between the transformed similarity graphs $S^{*(v)}$s is minimized across each view. Such transformations can be learnt by different approaches such as Min-Disagreement (de Sa, 2005), co-training (Kumar and Daumé, 2011) or co-regularization (Kumar et al., 2011). Finally, standard spectral clustering (Ng et al., 2002) can be applied to the transformed graph of the most informative view to obtain the final clustering result. This class of algorithms is related to the CoKmLDA in the sense that both approaches aim to identify clusters which are in consensus across each view, such that pairs of samples which are considered similar in one view should be considered similar in other views. However, the disadvantage of this class of algorithms is that, features $X^{(v)}$ are not used again after the $S^{(v)}$ is built. In the case that original features $X^{(v)}$ contain substantial number of noisy dimensions which are irrelevant to underlying classes, the estimation of $S^{(v)}$ is intrinsically inaccurate, thus improvements from graph fusion can be sub-optimal.

The second category of clustering approaches based on Canonical Correlation Analysis (CCA), i.e. (Chaudhuri et al., 2009; Blaschko and Lampert, 2008) aim to cope with multi-view, high-dimensional data. As illustrated in Fig. 2(b), the general idea involves jointly learning projections $P^{(1)}$ and $P^{(2)}$ with $X^{(1)}$ and $X^{(2)}$ such that the correlation between the projected samples in two views are maximized. Standard clustering algorithms such as k-means can then be applied to projected samples. The objective function is formulated as:

$$\arg\max_{P^{(1)},P^{(2)}} \frac{\mathbf{E}(P^{(1)}X^{(1)})\mathbf{E}(P^{(2)}X^{(2)})}{\sqrt{\mathbf{E}(P^{(1)}X^{(1)})^2\mathbf{E}(P^{(2)}X^{(2)})^2}} \tag{17}$$

We foresee two disadvantages of CCA based algorithms. First, according to the analysis of Chaudhuri et al. (2009), CCA learns a low dimensional subspace spanned by the means of different clusters (equivalent to the maximization of S_b). However, same cluster samples are not necessarily projected near to each other (minimiza-

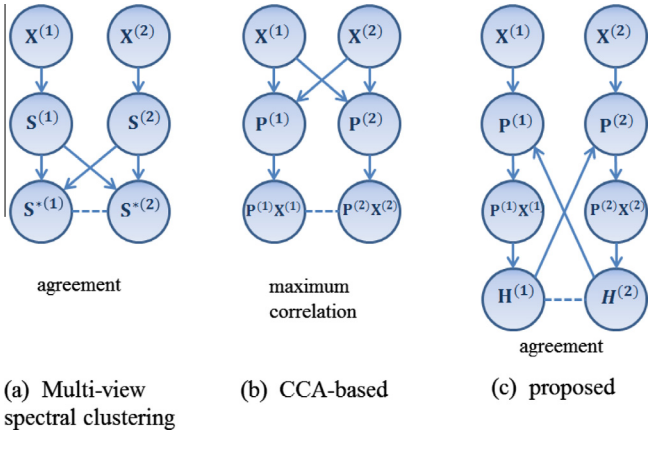


Fig. 2. Working flowchart of different multi-view clustering algorithms.

tion of S_w). Second, CCA-based methods rely strongly on the conditional independence assumption, which may not hold in practical problems. According the experimental work of Kumar and Daumé (2011) and Kumar et al. (2011), CCA-based method performs poorly when there is some dependence between views; this can be expected from Eq. (17). In the worst case, if $X^{(1)}$ and $X^{(2)}$ are fully correlated ($X^{(1)} = \alpha X^{(2)}$), any projections $P^{(1)} = P^{(2)}$ will maximize the objective function to its maximum value 1.

The framework of proposed CoKmLDA algorithm is illustrated in Fig. 2(c). CoKmLDA requires a maximum agreement between clustering results $H^{(1)}$ and $H^{(2)}$ on projected views $P^{(1)}X^{(1)}$ and $P^{(2)}X^{(2)}$. Compared to graph-based multi-view clustering algorithms, CoKmLDA reduces noise existed in features $X^{(1)}$ and $X^{(2)}$ through the iterative learning of projections $P^{(1)}$ and $P^{(2)}$ whereas graph-based methods reduce noises in similarity graphs. Compared to CCA-based multi-view clustering algorithms, CoKmLDA directly requires maximum agreement of clustering results rather than maximum correlation in projected spaces. Moreover, CoKmLDA is less sensitive to the view dependency. After all, CoKmLDA algorithm is equivalent to single-subspace clustering algorithm LDA-km proposed in Ding and Li (2007). Finally, as we will shown in Section 5.3, CoKmLDA can exploit the existed independence between views even if it is weak.

5. Experiments and discussions

In this section, we evaluate the effectiveness of the proposed algorithm on 3 independent datasets under 2 scenarios, when the conditional independence assumption is fully satisfied or only partially satisfied. For the former, its performance is assessed with audio-visual person clustering based on facial and speech features on the MOBIO database [1](McCool et al., 2012). For the later, we report its application to image clustering using the UCI handwritten digit dataset,[2] and text document clustering using BBC News Synthetic multi-view text dataset.[3] Note that the complexity of the proposed CoKmLDA algorithm grows linearly with the square of feature length (as discussed in Section 3.1), so for the efficiency of computation, in all experiments, all features are reduced to 100 dimensions by a PCA preprocessing step. All results are averaged by across independent trials of random initialization.

The performance of CoKmLDA is compared to four baseline systems: conventional k-means in PCA space, the LDA-km single-view subspace clustering algorithm (Ding and Li, 2007) and two other recently proposed multi-view clustering algorithms, namely Canonical Correlation Analysis (CCA) (Chaudhuri et al., 2009) and co-training spectral clustering (CoSC) (Kumar and Daumé, 2011). These latter two algorithms represents the two different approaches to multi-view clustering algorithms discussed in Section 4. Despite some underlying limitations of the component algorithms of the proposed approach, e.g., LDA is only optimal for Gaussian-distributed data and k-means often gives unreliable clustering of non-spherical-shaped data, experiments with three real-world datasets presented in this section show that the proposed CoKmLDA algorithm gives satisfactory clustering performance in solving practical problems.

5.1. Evaluation metrics

The clustering performances of the proposed CoKmLDA algorithm and other compared methods are assessed using four different metrics, namely clustering accuracy, Normalized Mutual Information (NMI), and pairwise precision and recall. The clustering accuracy is given by:

$$CA = \frac{1}{n}\sum_{i=1}^{n}\delta(h_i, \mathrm{map}(l_i)) \tag{18}$$

where n is the number of samples, h_i is the cluster indicator estimated for the ith sample, l_i is the corresponding true label, and $\delta(a,b)$ is a function which returns 1 if $a = b$ and 0 otherwise. The $map(\,)$ function represents the mapping between cluster indicators and true labels as determined according to a Hungarian algorithm (Steiglitz and Papadimitriou, 1982).

The normalized mutual information (NMI) (Strehl and Ghosh, 2003) is another popular clustering metric derived from information theory and given by:

$$NMI = \frac{I(H,L)}{\sqrt{E(H)E(L)}} \tag{19}$$

where $I(H,L)$ is the mutual information between H and L and $E(H)$ and $E(L)$ are the respective entropy. The NMI lies between 0 and 1 and larger values indicate more accurate clustering indicators. Please see (Strehl and Ghosh, 2003) for more details.

The pairwise precision and recall are defined in the following way. Let true positives (TP) be the number of pairs of same-class samples assigned to the same cluster, false positives (FP) be the number of pairs of different-class samples assigned to the same cluster, and false negatives (FN) be the number of pairs of same-class samples assigned to different clusters. The precision and recall are then given by:

$$Precision = \frac{TP}{TP+FP}; \quad Recall = \frac{TP}{TP+FN} \tag{20}$$

Both metrics take value between 0 and 1.

5.2. Audio-visual speaker clustering (conditional independence assumption satisfied)

We first evaluate the effectiveness of the proposed algorithm through experiments in audio-visual speaker clustering. In this case, each view is conditionally independent and represented with features of high dimensionality. Facial features are corrupted by inter-session variations such as illumination, expression and pose whereas vocal features are corrupted by different phonemes pronounced in a short speech episode, which are expected to be independent from each other.

5.2.1. Database and feature extraction

We consider speaker clustering using speech and facial images. Experiments are conducted with the standard MOBIO database (McCool et al., 2012) which contains videos of 150 subjects captured in real-world, challenging conditions. Recordings come from a mobile phone camera and are captured in 12 different sessions over a 18-month period where each session contains 11–21 videos. Fig. 3 illustrates the level of inter-session variation in a set of example frames for a given subject. For computational efficiency, we test our algorithm using a subset of data from 40 male subjects and for each of them, 5 videos are selected from each of the 12 sessions. This results in a pool of 2400 video samples.

We use cropped face images provided with the MOBIO database, one image per video sample. All images are resized to 50×43 pixels and then histogram equalized. Rows of pixel intensities are concatenated to form feature vectors of 2150 dimensions. The speech signal is split into frames of 20 ms duration before the extraction of features composed of 26 Mel-scaled frequency cepstral coefficients (MFCCs), their 26 derivatives and the delta energy. Energy-based voice activity detection is then applied to discard

[1] https://www.idiap.ch/dataset/mobio
[2] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[3] http://mlg.ucd.i.e./datasets/segment.html

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

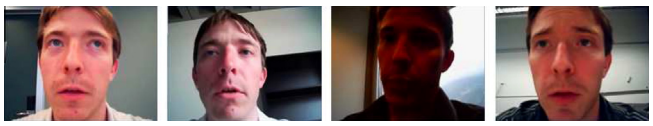


Fig. 3. Sample facial images of a subject in different sessions in MOBIO database.

non-speech frames. A 64-component Gaussian mixture model (GMM) is then fitted to remaining speech data through the maximum a posteriori (MAP) adaptation of a speaker-independent world model. The means of the GMM model are then concatenated to form a 3392-dimensional GMM supervector (Reynolds et al., 2000). Both face and speech feature vectors are first reduced to 100 dimensions through the application of PCA.

5.2.2. Results

The proposed CoKmLDA algorithm and all compared methods require the expected number of clusters K as an input parameter, which is set to be the number of subjects. In our experiments we observed that, for all linear subspace methods (PCA, LDA-km, CCA and CoKmLDA), the use of cosine-distance-based spherical k-means (Dhillon et al., 2001) consistently out-performs Euclidean-distance-based k-means. As a result, we report the results obtained with spherical k-means in these methods whereas for CoSC, we report the results obtained with conventional k-means which achieves the best performance in this case.

Table 2 summarizes the mean of the clustering accuracy, NMI, pairwise precision and pairwise recall obtained with 20 different runs of k-means with random initialization. It is observed that, for all metrics, multi-view clustering methods CCA, CoSC and CoKmLDA perform significantly better than the PCA baseline, whereas the single view LDA-km method only gives modest improvements over the PCA baseline. Finally, the proposed CoKmLDA algorithm outperforms the closest-performing method CCA by a significant margin (over 10% gain in clustering accuracy and approximately 5% in NMI). Fig. 4 shows the variation in accuracy and CAI scores as a function of the number of iterations. Convergence is seen to occur in fewer than 15 iterations. In practice we have not encountered any cases where convergence does not occur.

5.2.3. Clustering visualisations and discussion

All the approaches compared above embed data samples into lower dimensional spaces in which clustering is then performed with a standard k-means algorithm. PCA, LDA-km, CCA, and the proposed CoKmLDA algorithm embed original data into linear subspaces, while co-training spectral clustering embeds data samples into the first K eigenvectors of the graph Laplacian (Kumar and Daumé, 2011). It is informative to visualize the embedded data structure and thus to observe the relationship between the embedded structure and clustering performance. However, the embedded subspaces are still high-dimensional and cannot be visualized directly. T-distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten and Hinton, 2008) is a powerful tool used to visualize high-dimensional data via the embedding of data into a 2-D or 3-D space while respecting relative distances between data samples.

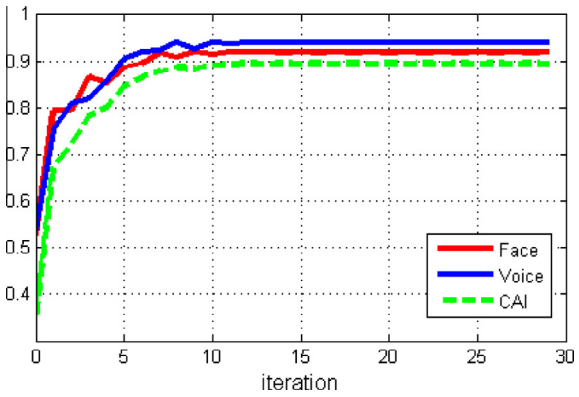


Fig. 4. Clustering accuracy for face modality (red), voice modality (blue) and CAI score (green) v.s. number of iterations of co-training. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Performance comparison.

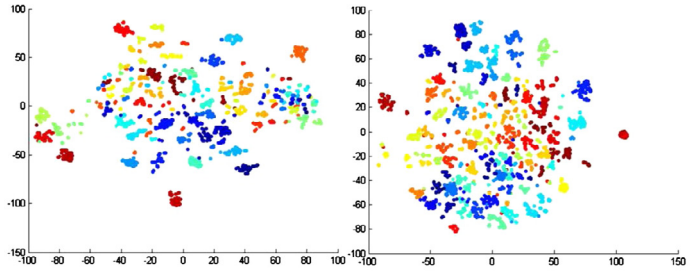
	Face		Speech	
	Accuracy	NMI	Accuracy	NMI
(a) Mean clustering accuracy and NMI score over 20 random trials				
PCA	0,530	0,665	0,512	0,667
LDA-Km (Ding and Li, 2007)	0,712	0,842	0,668	0,815
CCA (Chaudhuri et al., 2009)	0,825	0,915	0,798	0,924
CoSC (Kumar and Daumé, 2011)	0,785	0,895	0,799	0,895
CoKmLDA	**0,934**	**0,970**	**0,910**	**0,959**
	Face		Speech	
	Precision	Recall	Precision	Recall
(b) Mean precision and recall over 20 random trials				
PCA	0,359	0,396	0,425	0,407
LDA-Km (Ding and Li, 2007)	0,652	0,765	0,585	0,692
CCA (Chaudhuri et al., 2009)	0,675	0,882	0,677	0,864
CoSC (Kumar and Daumé, 2011)	0,662	0,861	0,652	0,841
CoKmLDA	**0,889**	**0,972**	**0,864**	**0,952**

Fig. 5 illustrates 2-D scatter plots of projected data for PCA, LDA-km, CCA, CoSC and CoKmLDA after the application of t-SNE. In all cases, samples belonging to different classes are represented by different colors. The features processed by PCA is shown in Fig. 5(a). The sample distribution is especially noisy which explains the poor clustering performance. In Fig. 5(b), clearer cluster structures are observed in LDA-km subspaces but the confusion between several classes is still high, due to its single-view nature. In CCA subspaces (Fig. 5(c)), cluster structure is not visually obvious. Same-class samples are approximately located in one single Gaussian distribution, but the variance is relatively high, since CCA does not minimize with-in class scatter, as discussed in Section 4. Both CoSC and the proposed CoKmLDA produce large between-class/within-class scatter ratio, as shown in Fig. 5(d) and (e) respectively. However, the clustering purity in CoKmLDA subspaces is significantly better.

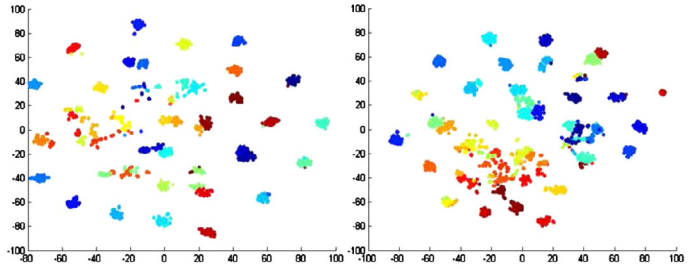
In the following we address some potential anomalies in the reported results. Fig. 5 and Table 2 show that, while CoSC gives better cluster separation, performance is worse than that of CCA. Even though CoSC produces a subspace in which different clusters are better separated, the data structure produced with CCA is cleaner with respect to the true labels. However, with a better separated cluster structure, more sophisticated initialization method for the k-means algorithm may deliver improved clustering performance.

It is also of interest to reflect on the reasons why CoKmLDA delivers such significantly better performance than other approaches. We attribute the superior performance of CoKmLDA to two main factors. First, CoKmLDA learns discriminative subspaces in which the cluster structure is in agreement for each view. Insodoing, the influence of feature dimensions which are unrepresentative of the underlying class structure is greatly reduced. Second, as discussed in Section 3.1, seeds used for k-means in each iteration are inherited from samples closest to the K centroids identified

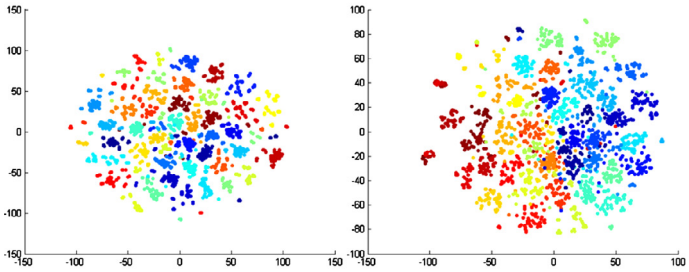
Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003



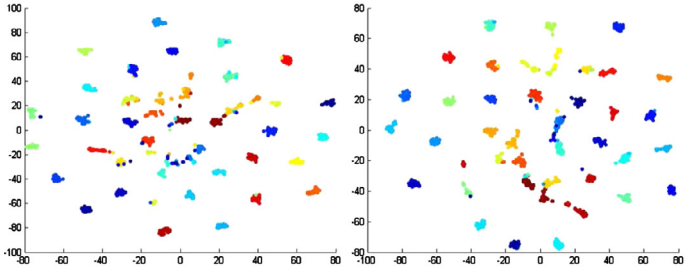
(a) PCA embeddings for face(left) and voice(right)



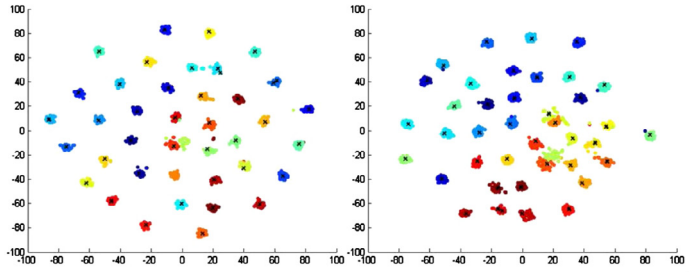
(b) LDA-Km embeddings for face(left) and voice (right)



(c) CCA embeddings for face(left) and voice (right)



(d) CoSC embeddings for face(left) and voice (right)



(e) CoKmLDA embeddings face(left) and voice (right).
Black crosses represents k-mean seeds inherited

Fig. 5. 2-D t-SNE visualizations of data structures for PCA, CCA, CoSC, and CoKmLDA subspaces. Different subjects/classes are represented by different colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

Table 3
Performance comparison on UCI Handwritten digits dataset.

	View 1 (Fou)		View 2 (Fac)		View 3 (Pix)	
	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
(a) Mean clustering accuracy and NMI score over 20 random trials. Number (2) or (3) indicates the number of views used in the approach.						
PCA	0,525	0,603	0,603	0,651	0,601	0,642
LDA-Km (Ding and Li, 2007)	0,576	0,677	0,750	0,798	0,771	0,804
CCA (Chaudhuri et al., 2009)	0,542	0,647	0,644	0,687	N.A.	
CoSC(2) (Kumar and Daumé, 2011)	0,702	0,752	0,748	0,774	N.A.	
CoKmLDA(2)	**0,725**	**0,769**	**0,761**	**0,810**	N.A.	
CoSC(3)(Kumar and Daumé, 2011)	**0,740**	**0,773**	0,772	0,793	0,764	0,782
CoKmLDA(3)	0,720	0,759	**0,892**	**0,852**	**0,845**	**0,844**
	View 1 (Fou)		View 2 (Fac)		View 3 (Pix)	
	Precision	Recall	Precision	Recall	Precision	Recall
(b) Mean precision and recall over 20 random trials. Number (2) or (3) indicates the number of views used in the approach.						
PCA	0,525	0,594	0,568	0,601	0,593	0,620
LDA-Km(Ding and Li, 2007)	0,527	0,615	0,702	0,759	0,642	0,739
CCA(Chaudhuri et al., 2009)	0,568	0,620	0,604	0,691	N.A.	
CoSC(2)(Kumar and Daumé, 2011)	0,665	0,685	0,698	0,757	N.A.	
CoKmLDA(2)	**0,685**	**0,713**	**0,718**	**0,785**	N.A.	
CoSC(3)(Kumar and Daumé, 2011)	**0,704**	**0,730**	0,756	0,798	0,744	0,787
CoKmLDA(3)	0,675	0,701	**0,821**	**0,860**	**0,802**	**0,841**

in the preceding iteration and the algorithm tends to give one seed per compact cluster. This fact is shown in Fig. 5(e), where the black crosses represents the seeds of k-means automatically learnt by CoKmLDA algorithm.

5.3. Handwritten digit clustering and text document clustering (conditional independence assumption not fully satisfied)

Co-training-style algorithms generally assume the conditional independence between the multiple features in use. However, in many practical problems, this assumption is not fully validated. As opposed to different features from different sources (as with visual and audio sources in the previous example), when both features come from the same source, they are expected to be correlated to some extent. To assess the CoKmLDA algorithm in such settings, we report further experiments with the clustering of image-only and text-only documents.

5.3.1. Databases

The proposed algorithm is assessed using two different databases: the UCI handwritten digits dataset[4] for image clustering with different features, and the BBC News Synthetic multi-view text dataset[5] for text document clustering.

The UCI handwritten digits dataset consists of images of handwritten numerals ('0'–'9') extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. The database provides multiple type of features extracted from the images. We used the 76-dimensional Fourier coefficients (Fou) as view 1 and the 216-dimensional profile correlations (Fac) as view 2. In order to assess the effectiveness of the extension of CoKmLDA to more than two views, we further choose the pixel intensities (Pix) as the third view, which is a 240-dimensional feature vector. Note that the first view is intrinsically less informative than the two others since the Fourier coefficients are rotation invariant hence cannot distinguish between the digit '6' and the digit '9'. Both profile correlation and pix intensity features are reduced to 100 dimensions by PCA as in audio-visual person clustering experiment, while for Fourier coefficient features are only pre-processed by removing the mean since it has only 76 dimensions.

The BBC News synthetic multi-view text dataset consist of term frequency features from news articles from the BBC (Greene and Cunningham, 2005). BBC data contains 2225 complete news articles corresponding to stories in five topical areas (business, entertainment, politics, sport and technology). Each document is segmented into two parts, and word frequency features are computed from each part, which constitute the two views (seg1of2 & seg2of2) (Greene and Cunningham, 2009). The feature dimension is 6838 and 6790 for the two views, respectively. Both features are reduced to 100 dimensions by PCA.

5.3.2. Results and discussions

Clustering performance is again assessed in terms of clustering accuracy, NMI, pairwise precision and recall. Table 3 shows results for the UCI handwritten dataset. The number in the parentheses (2 or 3) after CoSC and CoKmLDA indicates the number of views used in the algorithm. Since Chaudhuri et al. (2009) does not provide an extension to more than two views for CCA-based clustering, in this case results are reported for the first two views only. According to the most informative view (Fac), The proposed CoKmLDA algorithm gives the best performance among all methods compared. For the two-view setting, and the most informative view (Fac), the single-view LDA-km algorithms performs closest to the CoKmLDA algorithm. However, CoKmLDA is still successful in utilizing information in the first view (Fou), even if it is less informative, and performs marginally better than the single-view algorithm. This observation shows that the proposed algorithm can exploit even-marginal independence between views. CCA only provides marginal improvements over the PCA baseline, due to its sensitivity to correlated features. When the additional third view is used, the CoKmLDA algorithm gives a further 12% increase in terms clustering accuracy for the most informative view (Fac), which demonstrates the effectiveness of the multi-view extension proposed in Section 3.4.

Table 4 summarizes results for BBC News dataset. The proposed algorithm still out-performs all the compared methods and the co-spectral-clustering algorithm is the second-best performing algorithm. Note that the CCA method performs even worse than the PCA baseline. These results confirm the analysis in Section 4

[4] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[5] http://mlg.ucd.i.e./datasets/segment.html

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003

Table 4
Performance comparison on BBC News dataset.

	View 1 (seg1of2)		View 2 (Seg2of2)	
	Accuracy	NMI	Accuracy	NMI
(a) Mean clustering accuracy and NMI score over 20 random trials				
PCA	0,852	0,701	0,863	0,713
LDA-Km (Ding and Li, 2007)	0,877	0,762	0,882	0,755
CCA (Chaudhuri et al., 2009)	0,725	0,688	0,746	0,692
CoSC (Kumar and Daumé, 2011)	0,886	0,762	0,887	0,775
CoKmLDA	**0,912**	**0,788**	**0,915**	**0,803**
	View 1 (seg1of2)		View 2 (Seg2of2)	
	Precision	Recall	Precision	Recall
(b) Mean precision and recall over 20 random trials				
PCA	0,795	0,802	0,801	0,812
LDA-Km (Ding and Li, 2007)	0,794	0,806	0,804	0,821
CCA (Chaudhuri et al., 2009)	0,678	0,797	0,710	0,801
CoSC (Kumar and Daumé, 2011)	0,833	0,845	0,841	0,849
CoKmLDA	**0,852**	**0,861**	**0,858**	**0,867**

that CCA method strongly relies on the assumption of conditional independence between views and is risky to use when this assumption no longer holds. The proposed CoKmLDA algorithm, on the other hand, is more reliable when the conditional independence assumption is weak.

6. Conclusions

This paper proposes a new co-training framework for unsupervised, multi-view subspace clustering. It applies the results of unsupervised clustering in one view to learn discriminant subspaces in another. The general framework assumes conditionally independent views. We show, however, that the new algorithm still performs well when the conditional independence is weak. Furthermore, the framework is straightforward and combines well-known, even trivial algorithms to positive effect. The paper also presents a theoretical treatment which shows how LDA projections learned from samples with random label noise are equivalent to those learned with entirely clean labels and that the cross-view labeling, or co-training, is efficient in correcting erroneous sample labels. Experiments in audio-visual speaker clustering, multi-view handwritten digit clustering and text document clustering demonstrate the effectiveness of our algorithm and superior performance to existing state-of-the-art approaches.

References

Basu, S., Banerjee, A., Mooney, R., 2002. Semi-supervised clustering by seeding. In: Proceedings of 19th International Conference on Machine Learning (ICML).
Baudat, G., Anouar, F., 2000. Generalized discriminant analysis using a kernel approach. Neural Comput. 12 (10), 2385–2404.
Bickel, S., Scheffer, T., 2004. Multi-view clustering. In: Proceedings of the IEEE International Conference on Data Mining. vol. 36.
Blaschko, M.B., Lampert, C., 2008. Correlational spectral clustering, In: CVPR 08.
Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM, pp. 92–100.
Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K., 2009. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 129–136.
de Sa, V.R., 2005. Spectral clustering with two views. In: Proceedings of Workshop of Learning with Multiple Views.
Dhillon, I.S., Modha, D.S., 2001. Concept decompositions for large sparse text data using clustering. Mach. Learn. 42 (1–2), 143–175, URL <http://dx.doi.org/10.1023/A:1007612920971>.
Ding, C., Li, T., 2007. Adaptive dimension reduction using discriminant analysis and k-means clustering. In: International Conference on Machine Learning. Academic Press, pp. 84–405.
Greene, D., Cunningham, P., 2005. Producing accurate interpretable clusters from high-dimensional data. Knowledge Discovery Databases PKDD 2005, 486–494.
Greene, D., Cunningham, P., 2009. A matrix factorization approach for integrating multiple data views. Mach. Learn. Knowledge Discovery Database, 423–438.
Jolliffe, I., 2005. Principal Component Analysis. Wiley Online Library.
Kriegel, H.-P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowledge Discovery Data (TKDD) 3 (1), 1–58.
Kumar, A., Daumé III, H., 2011. A co-training approach for multi-view spectral clustering. In: International Conference on Machine Learning.
Kumar, A., Rai, P., Daumé III, H., 2011. Co-regularized multi-view spectral clustering. Adv. Neural Inf. Process. Syst. 24, 1413–1421.
McCool, C., Marcel, S., Hadid, A., Pietikainen, M., Matejka, P., Cernocky, J., Poh, N., Kittler, J., Larcher, A., Levy, C., Matrouf, D., Bonastre, J., Tresadern, P., Cootes, T., 2012. Bi-modal person recognition on a mobile phone: using mobile phone data. In: 2012 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 635–640.
Ng, A.Y., Jordan, M.I., Weiss, Y., et al., 2002. On spectral clustering: analysis and an algorithm. Adv. neural inf. process. syst. 2, 849–856.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. In: Digital Signal Processing. p. 2000.
Steiglitz, K., Papadimitriou, C.H., 1982. Combinatorial Optimization: Algorithms and Complexity. Prentice Hall, New Jersey, UV Vazirani (1984). On two geometric problems related to the travelling salesman problem. J. Algorithms 5, 231–246.
Strehl, A., Ghosh, J., 2003. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617.
Tang, H., Chu, S., Hasegawa-Johnson, M., Huang, T., 2012. Partially supervised speaker clustering. IEEE Trans. Pattern Anal. Mach. Intell. 34 (5), 959–971, URL <http://dx.doi.org/10.1109/TPAMI.2011.174b0110>.
van der Maaten, L., Hinton, G., 2008. Visualizing high-dimensional data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.
Wang, H., Ding, C., Huang, H., 2010. Multi-label linear discriminant analysis. In: Computer Vision – ECCV 2010. Springer, pp. 126–139.
Ye, J., Zhao, Z., Wu, M., 2007. Discriminative k-means for clustering. Adv. Neural Inf. Process. Sys. 20, 1649–1656.

Please cite this article in press as: Zhao, X., et al. A subspace co-training framework for multi-view clustering. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.12.003