Pattern Recognition Letters 41 (2014) 73–82

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 8 December 2013

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: