Pattern Recognition Letters 77 (2016) 35–41

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Non-negative embedding for fully unsupervised domain adaptation☆

Ievgen Redko*, Younès Bennani

LIPN, CNRS (UMR 7030), Université Paris 13, Sorbonne Paris Cité, 99, avenue Jean-Baptiste Clment, Villetaneuse, 93430, France

A R T I C L E I N F O

Article history:
Received 21 September 2015
Available online 26 March 2016

Keywords:
Transfer learning
Domain adaptation
Non-negative matrix factorization

A B S T R A C T

Domain adaptation is a field of machine learning that addresses the problem occurring when a classifier is trained and tested on domains from different distributions. This kind of paradigm is of vital importance as it allows a learner to generalize the knowledge across different tasks. In this paper, we present a new method for fully unsupervised domain adaptation that seeks to align two domains using a shared set of basis vectors derived from eigenvectors of each domain. We use non-negative matrix factorization (NMF) techniques to generate a non-negative embedding that minimizes the distance between projections of source and target data. We present a theoretical justification for our approach by showing the consistency of the similarity function defined using the obtained embedding. We also prove a theorem that relates the source and target domain errors using kernel embeddings of distribution functions. We validate our approach on benchmark data sets and show that it outperforms several state-of-art domain adaptation methods.

1. Introduction

Machine learning and data mining have already shown significant success in many areas of knowledge engineering including classification, regression and clustering. In the last decade, they have reached the level of accuracy of a human being in many real-world classification tasks, including but not limited to automatic speech recognition, image classification and natural language processing. These methods, however, work well only under a common assumption: the training and test data are drawn from the same feature space and from the same distribution. When the distribution changes, most statistical models must be rebuilt from newly collected data. In many real-world applications, it is expensive or impossible to obtain new data needed to reconstruct the learning models. Therefore, it is necessary to develop approaches that reduce the need and the effort of collecting new data by aligning domains while supposing that there exists a good classifier for both of them. Algorithms that tackle this problem are usually called domain adaptation approaches. For example, applying a classifier trained on the images from Amazon online merchants to web camera photos could be beneficial in case the shift between domains has been taken into account. Domain adaptation involves two interrelated problems, aiming at learning a robust classifier in source domain hoping that it will perform well in the related target domain by reducing the discrepancy between their distributions.

There are two main categories for domain adaptation algorithms: semi-supervised domain adaptation (all source domain data are labeled and only a few data instances are labeled in target domain) and unsupervised domain adaptation (all source data are labeled and no labeled data are available for the target domain). Last scenario is a topic of ongoing interest among researchers as it reflects what actually happens when a system trained under perfect conditions on preprocessed data faces reality.

According to a survey on domain adaptation [20] there are four main classes of approaches that differ in the assumptions they make: instance weighting for covariate shift (matching reweighting instances between domains); selflabeling methods (guessing labels in target domain and adjusting them during the learning procedure); cluster-based learning (attributing the same label to the instances from dense regions); and feature representation learning (projecting initial features to a new space in order to find invariant components).

Last class draws considerably more attention in machine learning community due to the recent success of representation learning [2]. These methods usually proceed in one of two following ways: either they maximize the similarity between the distributions [5,22,23] or they try to find a shared latent feature space for both domains [3,15].

In our work, we would like to make use of the fact that some data are intrinsically non-negative and to show that preserving this non-negativity can be beneficial for classification while learning

☆ This paper has been recommended for acceptance by Dr. J. Laaksonen.

* Corresponding author. Tel.: +33 762 07 0967.
E-mail address: ievgen.redko@lipn.univ-paris13.fr (I. Redko).

http://dx.doi.org/10.1016/j.patrec.2016.03.016
0167-8655/

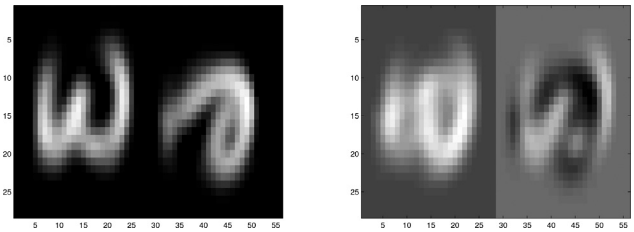

Fig. 1. Prototypes found by NMF (on the left) and PCA (on the right).

new feature representations. Indeed, it makes sense as almost all domain adaptation algorithms are usually applied to image classification and object recognition data sets where data represented by color frequencies are naturally non-negative. In Fig. 1 we visualize the prototypes obtained with non-negative constraints and without them on a subset of MNIST data set consisting of digits 3 and 6. We can note that the prototypes found by Non-negative matrix factorization (NMF)[17] are easily interpretable while those found by PCA are not. Finally, we also note that some preprocessing procedures used to obtain image descriptors may lead to negative values. They include Fisher Vectors, Vector of Locally Aggregated Descriptors (VLAD) and Deep Convolutional Activation Feature (DECAF). On the other hand, a vast majority of other descriptors, like Scale-Invariant Feature Transform (SIFT), Dense SIFT, Local Intensity Order Pattern (LIOP), Histogram of Oriented Gradients (HOG) and all covariant feature detectors with different operators preserve non-negativity and can be used as input matrices for our approach.

The idea of preserving non-negativity has already been used with success by a couple of approaches for transfer learning and domain adaptation through NMF techniques, notably Tri-NMF [6,8,18,30].

First, our approach obtains a reduced set of non-negative principal components of the source domain where its cardinality is defined by the approach proposed in [11]. Then we simultaneously learn a shared dictionary for both domains using a simple cost function and apply it further in the target domain. These two steps allow us to preserve the non-negativity of data all along the learning procedure and find a non-negative embedding for both domains as shared basis vectors between two sets of correlated non-negative principal components.

We also show how a theorem that relates source and target domains from classical domain adaptation theory can be restated using kernel embeddings of distribution functions. This allows us to provide an efficient way to measure the discrepancy between distributions using a divergence measure which quadratic approximation can be estimated in linear time.

The rest of our paper is organized as follows: in Section 2 we describe in more detail the key differences between our approach and other similar approaches; in Section 3 we present our method and prove a consistency theorem for similarity function based on nonnegative embedding; in Section 4 we give a simple generalization of the theorem that relates source and target domain errors; in Section 5 we derive the multiplicative update rules for our optimization problem. Finally, we evaluate our approach in Section 6 and end by conclusions in Section 7.

2. Related works

Our work is related to a couple of unsupervised domain adaptation methods where the goal is to find an intermediate representation of source domain data that can be used further in combination with a target domain. A common approach is to look for a new projection of data in the corresponding space. To this end, the application of PCA was widely investigated and used in order to find a common space where the divergence between marginal distributions of two domains is minimized [5,23]. According to the theory of domain adaptation [1] classification error of the target task is bounded by the divergence between distributions of each domain so this idea is theoretically justified. Subspace approaches were also widely used for domain adaptation and transfer learning in, for example, [9,11].

Our approach differs from the above mentioned works in two principal ways. First, we seek to find a non-negative embedding of basis vectors for two domains so that we could benefit from the fact that some data are intrinsically non-negative. To this end, our approach is similar to methods presented in [6,18,30]. The main difference, however, is that they use NMF techniques to match objects of two domains and they are usually applied for text-classification via Tri-NMF. Our approach learns a shared dictionary instead and its application is not limited to text classification. The second main difference is that we do not assume that we have labels in source domain – we use a shared set of basis vectors simultaneously with performing clustering of the target domain data. This type of setting is usually called “self-taught clustering” [7] and to the best of our knowledge there are no other NMF-based methods for domain adaptation that does not assume the presence of labels in the source domain. In our work, we would like to show that this paradigm can be used as a complementary to unsupervised domain adaptation approaches.

3. Unsupervised domain adaptation via non-negative embedding

We assume that we have two sets of unlabeled data $X_S \in \mathbb{R}^{m\times n_s}$ and $X_T \in \mathbb{R}^{m\times n_t}$ that correspond to source task data and target task data respectively. We denote their marginal distributions by $\mathcal{D}_S$ and $\mathcal{D}_T$. The first step of our approach consists in retrieving non-negative basis of each task by applying Projective NMF to X_S and X_T.

3.1. Projective NMF

Orthogonal Projective NMF (OPNMF) introduced in [28] minimizes the following cost function:

$$\min\ J = \|X - UU'X\|_F^2 \tag{1}$$

$$s.t.\ U'U = I, X, U \geq 0, \tag{2}$$

where

- $X \in \mathbb{R}^{m\times n}$ is an input data matrix;
- columns of $U \in \mathbb{R}^{m\times k}$ can be considered as basis vectors;
- k is the desired number of basis vectors.

Orthogonal Projective NMF solves PCA problem with non-negative constraints when Oja’s rule is applied during the optimization procedure.

As a first step we apply OPNMF to matrices X_S, X_T and we fix $k = d^*$:

$$X_S \simeq U_S U_S' X_S, X_S \in \mathbb{R}_+^{m\times n_s}, U_S \in \mathbb{R}_+^{m\times d^*}, \tag{3}$$

$$X_T \simeq U_T U_T' X_T, X_T \in \mathbb{R}_+^{m\times n_t}, U_T \in \mathbb{R}_+^{m\times d^*}. \tag{4}$$

The resulting matrices U_S and U_T are d^* non-negative principal components of X_S and X_T. We cannot use source principal components directly for target task as we are interested in features that are aligned with target task basis vectors. To choose them we use a slightly modified subspace disagreement measure (SDM) presented in [11] to define d^* nonnegative principal components for source task.