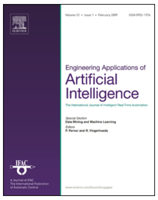
Engineering Applications of Artificial Intelligence 28 (2014) 181–189

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Beyond cross-domain learning: Multiple-domain nonnegative matrix factorization

Jim Jing-Yan Wang [a,b], Xin Gao [a,*]

[a] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

[b] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Article history:
Received 19 February 2013
Received in revised form
7 October 2013
Accepted 1 November 2013
Available online 12 December 2013

Keywords:
Data representation
Nonnegative matrix factorization
Cross-domain learning

ABSTRACT

Traditional cross-domain learning methods transfer learning from a source domain to a target domain. In this paper, we propose the multiple-domain learning problem for several equally treated domains. The multiple-domain learning problem assumes that samples from different domains have different distributions, but share the same feature and class label spaces. Each domain could be a target domain, while also be a source domain for other domains. A novel multiple-domain representation method is proposed for the multiple-domain learning problem. This method is based on nonnegative matrix factorization (NMF), and tries to learn a basis matrix and coding vectors for samples, so that the domain distribution mismatch among different domains will be reduced under an extended variation of the maximum mean discrepancy (MMD) criterion. The novel algorithm — multiple-domain NMF (MDNMF) — was evaluated on two challenging multiple-domain learning problems — multiple user spam email detection and multiple-domain glioma diagnosis. The effectiveness of the proposed algorithm is experimentally verified.

1. Introduction

The cross-domain learning problem has attracted much attention from the machine learning community recently. The goal of the problem is to learn a classifier to classify the samples in a target domain, but the number of labeled samples in the target domain is not sufficient for the learning. At the same time, it assumes that a source domain with a sufficient number of labeled samples exists and could be helpful for the learning problem of the target domain. The source domain samples share the same feature space and the class label space as the target domain, but the distribution of source domain samples is significantly different from that of the target domain. Thus the source domain samples cannot be directly used for the learning problem of the target domain. Instead, the domain transfer, or domain adaptation, is needed to learn from the source domain to the target domain. Recently, many works have been done in the transfer learning problem (Daume, 2007; Yang et al., 2007; Jiang et al., 2008; Bruzzone and Marconcini, 2010; Duan et al., 2012a,b).

Previous cross-domain learning research mainly focused on learning from a single source domain to a target domain (Daume, 2007; Yang et al., 2007; Jiang et al., 2008; Bruzzone and Marconcini, 2010; Duan et al., 2012a,b). It usually assumes that there are only a few labeled samples in the target domain, while a great number of labeled samples in the source domain. However, in many real-world applications, there are usually more than two domains, and for each domain there are just a few labeled samples. In such cases, the goal is to learn a classifier for each domain with the help of labeled samples from all other domains, i.e., each domain could be a target domain and at the same time, it can also be a source domain for all other domains. For example, in the problem of spam email detection, we may have several email subsets of different users. For each user, a large number of emails are collected while only a small portion of them are labeled as non-spam or spam. Because the data distributions of the different users' emails are different but related, they could be treated as different domains. Every user needs a classifier for the spam detection, but does not have enough labeled emails, thus every one is a target domain. At the same time, each user's data would also be helpful for the learning of other users' classifier, thus they are also target domains. We define this problem as the *multiple-domain learning* problem. It has the following features:

1. Several (usually more than two) domains will be considered in the learning procedure.
2. All the domains will be treated equally. No domain will be specified as source or target domains. Each domain could be

[*] Corresponding author. Tel.: +966 2 808 0323.
E-mail addresses: jimjywang@gmail.com (J.J.-Y. Wang), xin.gao@kaust.edu.sa (X. Gao).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.11.002

a target domain, while could also be a source domain for other domains.

3. A classifier is needed for each domain, but the number of labeled samples for each domain is limited.

Though there are many real-world applications of the multiple-domain learning problem, surprisingly, little attention has been paid in the learning problem of multiple-domains. Recently, a few methods have been presented to learn from multiple source domains to a single target domain. For example,

- Duan et al. (2012a) proposed the domain adaptation machine (DAM) for the multiple source domain adaption problem, which learns a robust classifier for the target domain by leveraging many base classifiers which could be learned using the labeled samples from the source domains or the target domains.
- Zhuang et al. (2010) proposed a centralized consensus regularization (CCR) framework for learning from multiple source domains to a target domain. It trains a local classifier by considering both local data available in one source domain and the prediction consensus with the classifiers learned from other source domains.
- Yao and Doretto (2010) proposed the multiple source transfer AdaBoost (MDTAB) by extending the boosting framework for transferring knowledge from multiple sources.
- Sun et al. (2011) proposed a two-stage-weighting-based method for multiple source domain adaptation (TSWMSD) (Sun et al., 2011), by combining weighted data samples from multiple sources based on marginal probability differences and conditional probability differences, with the target domain data.
- Tu and Sun (2012a) proposed a multiple source domain adaptation method based on ensemble learning. Using model-friendly classifiers, different test samples are assigned with different weights dynamically.
- Tu and Sun (2012b) further proposed a novel framework to maximize separability among classes and to minimize separability among domains simultaneously. To this end, the class-separate objectives and the domain-merge objectives are combined to achieve a unified objective.

All the aforementioned works tried to deal with the problem of learning from multiple source domains for a single target domain. They can be regarded as a special case of our multiple domain learning problem. The methods that learn from multiple source domains can be extended to the multiple-domain learning problem by treating each domain as the target domain and other domains as source domains in turn. However, this strategy has the following limitations:

- Its quite time consuming. For each domain, we should perform a learning procedure to learn form all other domains. It is not efficient when the number of domains is large.
- It assumes that all the samples in the source domains are labeled, which is not true for the multiple-domain learning problem. In this case, only the labeled samples from other domains will be utilized, while neglecting the unlabeled ones.
- The classifier is only learned for a specified target domain, and it could not be applied to other domains. Learning a single classifier for all the domains is impossible.

Nonnegative matrix factorization (NMF) has been well studied and applied as a data representation method, due to its ability to find the latent nature of data (Cai et al., 2011, 2009, 2008; Wang et al., 2012a, 2013a,b). However, up to now, all research carried on are limited in the single domain problem, and no cross-domain or multiple-domain NMF method has been studied yet. To fill these gaps, in this paper, we develop a novel data representation method for the multiple-domain data representation, based on NMF. We try to map all the samples from multiple-domains with different data distributions into a common representation space with a common distribution by NMF. A distribution mismatch term is constructed and applied to the coding vectors of samples under the framework of NMF. With the common representation of samples from multiple-domains, a robust classifier could be trained for the classification problem of samples of multiple-domains directly.

The rest of this paper is organized as follows: In Section 2, we propose the NMF learning algorithm for representation of samples from multiple-domains. In Section 3, we show the experimental results for the proposed algorithm on two real-world multiple-domain learning problems. Finally, we conclude this paper with the conclusion and possible future works in Section 4.

2. Multiple-domain nonnegative matrix factorization

In this section, we will introduce the proposed NMF method for representation of data samples from multiple-domains.

2.1. Objective function

To introduce the proposed NMF method, we construct an objective function for the factorization of all the samples from multiple-domains, by considering the following two problems simultaneously:

NMF problem: Given a training data set with N data samples denoted as $\mathcal{D}=\{x_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}_+^D$ is the D-dimensional nonnegative feature vector of the i-th sample, we organize it as a nonnegative matrix $X=[x_1,\dots,x_N]\in\mathbb{R}_+^{D\times N}$. NMF tries to find two low-rank nonnegative matrices $H\in\mathbb{R}_+^{D\times K}$ and $W\in\mathbb{R}_+^{K\times N}$ so that their product can approximate the original data matrix, $X\approx HW$ (Zheng et al., 2007; Gruber et al., 2009). $H=[h_1,\dots,h_K]$, where $h_k\in\mathbb{R}_+^D$ is the k-th column of H, could be regarded as a basis matrix with each column h_k being a basis vector. In this way, each sample x_i will be approximated as the linear combination of the basis vectors as

$$x_i \approx \sum_{k=1}^{K} h_k w_{ki} = Hw_i \tag{1}$$

where $w_i=[w_{1i},\dots,w_{Ki}]^\top$ is the linear combination coefficient vector for the i-th sample, and also the i-th column of W, which could be regarded as a new representation of x_i with regard to H. The coefficient matrix W is also called the coding matrix, and w_i is also called the coding vector of x_i.

To find the optimal factorization matrices H and W, we try to minimize the squared L_2 norm distance between the original matrix X and the product HW to make the approximation error as small as possible. The NMF problem can be formulated as the following constrained minimization problem:

$$\min_{H,W}\left\{\sum_{i:x_i\in\mathcal{D}}\|x_i-Hw_i\|_2^2=\|X-HW\|_2^2=\mathrm{Tr}[(X-HW)(X-HW)^\top]\right\}$$
$$\text{s.t.}\quad H\geq 0,\ \ W\geq 0. \tag{2}$$

where $\mathrm{Tr}(\cdot)$ is the trace of a matrix.

Multiple-domain distribution mismatch reduction problem: Suppose that the entire data set is composed of M domains