Engineering Applications of Artificial Intelligence 51 (2016) 163–170

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Optimal kernel choice for domain adaption learning

Le Dong [a,*], Ning Feng [a], Pinjie Quan [a], Gaipeng Kong [a], Xiuyuan Chen [a], Qianni Zhang [b]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), 2006 Xiyuan Avenue, Gaoxin West Zone, Chengdu, Sichuan 611731, China

[b] School of Electronic Engineering and Computer Science, Queen Mary, University of London, United Kingdom

ARTICLE INFO

Available online 2 March 2016

Keywords:
Optimal kernel
Domain adaption
Cross-domain
Test statistic
Kernel choice

ABSTRACT

In this paper, a kernel choice method is proposed for domain adaption, referred to as Optimal Kernel Choice Domain Adaption (OKCDA). It learns a robust classier and parameters associated with Multiple Kernel Learning side by side. Domain adaption kernel-based learning strategy has shown outstanding performance. It embeds two domains of different distributions, namely, the auxiliary and the target domains, into Hilbert Space, and exploits the labeled data from the source domain to train a robust kernel-based SVM classier for the target domain. We reduce the distributions mismatch by setting up a test statistic between the two domains based on the Maximum Mean Discrepancy (MMD) algorithm and minimize the Type II error, given an upper bound on error I. Simultaneously, we minimize the structural risk functional. In order to highlight the advantages of the proposed method, we tackle a text classification problem on 20 Newsgroups dataset and Email Spam dataset. The results demonstrate that our method exhibits outstanding performance.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Conventional machine learning methods universally assume that the training data and the test data come from the same distribution. Unfortunately for many applications, it is difficult to obtain enough labeled data for training classifiers. Recently, many researchers have been focusing on cross-domain adaption which aims at solving a learning problem in the target domain by utilizing training data in the source domain, while these two domains may have different distributions (Zhong et al., 2012; Pan et al., 2011). In practice, the domain adaptive learning strategy has been successfully applied to real-time applications, such as multi-task clustering (Zhang and Zhou, 2012), WiFi localization (Pan et al., 2008), action recognition (Wu et al., 2011), sentiment classification (Blitzer et al., 2007), visual event recognition (Duan et al., 2012a; Xu and Chang, 2008), object detection (Vzquez et al., 2011, 2014) and visual concept classification (Jiang et al., 2009; Yang et al., 2007; Jiang et al., 2008). However, compared with non-learning methods (Dong et al., 2012; Dong and Izquierdo, 2008), adaptive learning has more extensive applications.

To take the advantage of all labeled patterns for both auxiliary and target domains, Daume (2007) proposes a Feature Replication method to augment features for cross-domain learning. The augmented features are then used to construct a kernel function for Support Vector Machine training. Yang et al. (2007) propose Adaptive SVM for visual concept classification, in which the new classifier $f^T(x)$ is adapted from an existing classifier $f^A(x)$ trained from the source domain. Cross-domain SVM proposed by Jiang et al. (2009) uses k-nearest neighbors from the target domain to define a weight for each auxiliary pattern, and then the SVM classifier is trained with the re-weighted auxiliary patterns. More recently, Jiang et al. (2009) propose a method of mining the relationship among different visual concepts for video concept detection. They first build a semantic graph which can be adapted in an online fashion to fit the new knowledge mined from the test data. However, these methods do not utilize unlabeled patterns from the target domain. Such unlabeled patterns can also be used to improve the classification performance.

When there are only a few or even no labeled patterns available in the target domain, the auxiliary patterns or the unlabeled target patterns can be used to train the target classifier. Several cross-domain learning methods are proposed to cope with the inconsistency of data distributions. These methods re-weighted the training samples from the source domain by using unlabeled data from the target domain so that the statistics of samples from both domains are matched. Duan et al. (2012a,b) propose a cross-domain kernel learning framework, which learns a kernel function and classifier by minimizing both the structural risk functional and the distribution mismatch between the labeled and unlabeled

* Corresponding author. Tel.: +86 13981763623; fax: +86 28 61831655.
E-mail address: ledong@uestc.edu.cn (L. Dong).

http://dx.doi.org/10.1016/j.engappai.2016.01.022
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

samples from the auxiliary and target domains. This framework employs a domain similarity measure based on MMD. More recently, Duan et al. (2012a) develop a cross-domain learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL) that has been successfully used in visual event recognition.

A common insight is that most of those domain adaption learning methods are either variants of SVM or other kernel methods, which map auxiliary data and target data into a feature space for obtaining a robust SVM-based classifier, and simultaneously, minimize the mismatch between two different distribution domains. The performance of a classifier strongly depends on the choice of the kernels. Lanckriet et al. (2004) develop a nonparametric kernel matrix, which involves joint optimization of the coefficients in a conic combination of kernel matrices. One problem is that its time complexity is too high to be applied to real applications. In recent years, many effective methods (Duan et al., 2012b, c; Lu et al., 2014; Salah et al., 2014; Varma and Babu, 2009) have been developed to combine multiple kernels instead of directly learning the kernel matrix, in which the kernel function is a linear combination of based kernel functions. However, all those methods suppose that both test and training data are drawn from the same distribution. Consequently, naked multiple kernel learning cannot directly solve the problem of cross-domain learning. Because the coefficients of combination kernel are parameterized, the training data from source domain may degrade the performance of the model in the target domain.

In this paper, we propose a new method on kernel choice for cross-domain learning, which explicitly minimizes the loss due to the bias between the data distributions of the auxiliary and target domains, as well as the cost function of structural risk for all labeled patterns. Type I error is the probability of wrongly rejecting null hypothesis when the auxiliary distribution and the target distribution are drawn from the same distribution. Type II error is the probability of wrongly accepting null hypothesis when the auxiliary and the target distributions are different. Given an upper bound on Type I error, our kernel choice minimizes Type II error. The main contribution of this paper is that multiple base kernels are weighted to minimize the loss on the labeled examples and the bias between the data distributions in the two domains. Meanwhile, we minimize the bias between the source domain and the target domain by minimizing the Type II error. While multi-kernel method has been widely discussed (Bootkrajang and Kabán, 2014; Jia et al., 2014; Lu et al., 2014) and used (Salah et al., 2014), our work demonstrates that the kernel choice is pivotal to cross-domain learning.

The rest of paper is organized as follows: we briefly review the related works in Section 2. Section 3 introduces kernel choice for domain adaption learning. We experimentally compare the proposed method with other cross-domain learning methods on the 20 Newsgroups dataset and Email Spam dataset for text classification in Section 4. Finally, conclusion is made in Section 5.

2. Brief review of related work

Let us denote the dataset of labeled and unlabeled patterns from the target domain as $D_l^T=(x_i^T,y_i^T)|_{i=1}^{n_l}$ and $D_u^T=(x_i^T,y_i^T)|_{i=n_l+1}^{n_l+n_u}$, respectively, where y_i^T is the label of x_i^T, labeled patterns are numbered 1 to n_l, unlabeled patterns are numbered n_l+1 to n_l+n_u. We define $D^T=D_l^T\bigcup D_u^T$ as the dataset from the target domain with the size $n_t=n_l+n_u$ under the marginal data distribution ρ, and $D^A=(x_i^A,y_i^A)|_{i=1}^{n_A}$ as the dataset from the source domain under the marginal data distribution ϑ. We represent the labeled training dataset as $D=(x_i,y_i)|_{i=1}^{n}$, where n is the total number of labeled patterns. The labeled training data can be from the target domain ($D=D_l^T$) or from both domains ($D=D_l^T\bigcup D^A$).

2.1. Minimize bias of distribution using test statistic

It is important to reduce the mismatch between the source domain and the target domain distributions, and many methods have been proposed to address this work. A classic criteria is Kullback Leibler divergence (Rached et al., 2004). However, most of them are parametric and need to estimate an intermediate density. To steer clear of fussy measure, Borgwardt et al. (2006) present a novel non-parametric statistical method, namely, Maximum Mean Discrepancy, which is based on Reproducing Kernel Hilbert Space (Rosipal and Trejo, 2002):

$$\begin{aligned} MMD(D^A,D^T) &= \sup_{\|f\|_H\le 1}\left(E_{x^A\sim Q}[f(x^A)]-E_{x^T\sim P}[f(x^T)]\right)\\ &= \sup_{\|f\|_H\le 1}\langle f,(E_{x^A\sim Q}[f(x^A)]-E_{x^T\sim P}[f(x^T)])\\ &= \|E_{x^A\sim Q}[f(x^A)]-E_{x^T\sim P}[f(x^T)]\|_H, \end{aligned} \tag{1}$$

where $E_{x\sim\mu}[\cdot]$ denotes the expectation operator under the samples distribution μ and $f(x)$ is any function in H. The second equality holds as $f(x)=\langle f,\phi(x)\rangle_H$ by the property of RKHS, where $\phi(x)$ is the nonlinear feature mapping of the kernel k. Note that the inner product of $\phi(x_i)$ and $\phi(x_j)$ equals to the kernel function $k(\cdot,\cdot)$ on x_i and x_j, namely, $k(x_i,x_j)=\phi(x_i)\phi(x_j)$. An expression for the squared MMD is

$$\eta_k(D^A,D^T)=\|\phi(D^A)-\phi(D^T)\|_H^2 \tag{2}$$

$$\eta_k(D^A,D^T)=E_{xx'}k(x,x')+E_{yy'}k(y,y')-2E_{yx'}k(y,x'), \tag{3}$$

where $x,\ x'\sim^{i.i.d}p$ and $y,\ y'\sim^{i.i.d}q$. By introducing $h_k(x,x',y,y')=k(x,x')+k(y,y')-k(y,x')-k(x,y')$, Eq. (2) can be rewritten as $\eta_k=\mathbf{E}_{xx'yy'}h_k(x,x',y,y')$. By introducing $h_k(x,x',y,y')=k(x,x')+k(y,y')-k(y,x')-k(x,y')$, Eq. (2) can be rewritten as $\eta_k=\mathbf{E}_{xx'yy'}h_k(x,x',y,y')$. In brief, the key point of MMD is that the distance between distributions of two domains is equivalent to the distance between the means of the two domains mapped into a RKHS (Pan et al., 2008). Huang et al. (2006) develop a two-step method. The first step is to diminish the mismatch of means of different distributions in RKHS by reweighting the examples using square MMD. The second step is to learn a decision function that separates patterns from two opposite classes. One difficulty is that the performance of MMD strongly depends on the choice of kernel. Meanwhile, these methods do not ensure that the chosen kernel is optimal. Inspired by Gretton et al. (2012), we review the problem of bias between the source domain and the target domain as a two-sample test problem, which addresses the question of whether two independent samples are drawn from the same distribution. Consequently, given two example distributions: q from source (auxiliary) domain and p from target domain, we can set up a two-sample test which measures the similarity or bias between the source domain and the target domain.

We select some kernels for hypothesis testing from a particular family $\boldsymbol{K}$ of kernels, assuming kernel $k(x_i,x_j)$ is a linear combination of a set of base kernels

$$k_{\mathbf{d}}=\sum_{m=1}^{M}d_m k_m, \tag{4}$$

where $d_m>0$ is a set of positive coefficients, $\sum_{m=1}^{M}d_m=D>0$. The squared MMD becomes

$$\eta_m(D^A,D^T)=\|\phi(D^A)-\phi(D^A)\|_F^2=\sum_{l=1}^{M}d_l\eta_l(D^A,D^T). \tag{5}$$

Here, it is denoted that $d=\{d_1,d_2,\ldots,d_M\}^T\in R^{M\times 1}$, $\eta=\{\eta_1,\eta_2,\ldots,\eta_M\}\in R^{M\times 1}$. Eq. (5) can be written as $\eta_m(D^A,D^T)=d^T\eta$. η_m is the average of independent random variables, and its asymptotic distribution is given by the central limit theorem. Now we set up the

Download English Version:

https://daneshyari.com/en/article/380231

Download Persian Version:

https://daneshyari.com/article/380231

Daneshyari.com