Cross Domain Distribution Adaptation via Kernel Mapping

Erheng Zhong[†], Wei Fan[‡], Jing Peng[§], Kun Zhang[$],
Jiangtao Ren[†], Deepak Turaga[‡] and Olivier Verscheure[‡]
[†]Sun Yat-Sen University, Guangzhou, China
[‡]IBM T.J.Watson Research Center, New York, USA
[§]Montclair State University, New Jersey, USA
[$]Xavier University of Louisiana, Louisiana, USA
{sw04zheh@mail2, issrjt@mail}.sysu.edu.cn, {weifan, turaga, ov1}@us.ibm.com
pengj@mail.montclair.edu, kzhang@xula.edu

ABSTRACT

When labeled examples are limited and difficult to obtain, transfer learning employs knowledge from a source domain to improve learning accuracy in the target domain. However, the assumption made by existing approaches, that the marginal and conditional probabilities are directly related between source and target domains, has limited applicability in either the original space or its linear transformations. To solve this problem, we propose an adaptive kernel approach that maps the marginal distribution of target-domain and source-domain data into a common kernel space, and utilize a sample selection strategy to draw conditional probabilities between the two domains closer. We formally show that under the kernel-mapping space, the difference in distributions between the two domains is bounded; and the prediction error of the proposed approach can also be bounded. Experimental results demonstrate that the proposed method outperforms both traditional inductive classifiers and the state-of-the-art boosting-based transfer algorithms on most domains, including text categorization and web page ratings. In particular, it can achieve around 10% higher accuracy than other approaches for the text categorization problem. The source code and datasets are available from the authors.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – Data Mining

General Terms

Algorithms

1. INTRODUCTION

It is expensive or impractical for many applications to obtain large number of labeled examples. When this happens, most inductive learners perform poorly. The idea of transfer learning is to borrow labeled examples from a source-domain to improve learning in the target-domain. Let $P_t(\mathbf{x}, y)$ and $P_s(\mathbf{x}, y)$ denote joint distribution of target-domain and source-domain respectively. Supervised transfer learning is to use small number of labeled example from $P_t(\mathbf{x}, y)$, but many from $P_s(\mathbf{x}, y)$, to build a learning model for target-domain. The main challenge is to identify regions in $P(\mathbf{x}, y)$, either in original space or its transformation, where $P_t(\mathbf{x}, y)$ and $P_s(\mathbf{x}, y)$ are similar and knowledge can be transferred.

Much work has been proposed to solve transfer learning problem [15]. By definition, $P(\mathbf{x}, y) = r(y|\mathbf{x})q(\mathbf{x})$. Some work, such as [14], assumes that conditional probabilities $r_t(y|\mathbf{x})$ and $r_s(y|\mathbf{x})$ are similar in regions of the latent space where marginal distribution $q_t(\mathbf{x})$ and $q_s(\mathbf{x})$ of corresponding examples are close. Other works, such as [12], assumes $q(\mathbf{x})$ is related to $r(y|\mathbf{x})$. They both implicitly assume that marginal distribution and conditional probability are directly related. In summary, either of the following is assumed to be true and adopted to design transfer learning strategies: (1) where marginal distribution $q(\mathbf{x})$ of target-domain and source-domain are similar, conditional probability $r(y|\mathbf{x})$ also ought to be similar, or (2) vice versa. However, those assumptions may be too strict to be practical. For some problems, both the marginal and conditional distributions between target-domain and source-domain could be significantly different. When this happens, neither of the two assumptions is true anymore, in either the original space, scaled space or latent space using linear transformation.

However, non-linear transformation, such as the kernel manipulation, can make these assumptions plausible. A suitable kernel is able to map the input space into a convenient feature space where a linear boundary can be easily found [3]. Importantly, this also sheds light on transfer learning. First, a suitable kernel, such as Gaussian kernel [13], can make different input data form similar marginal distribution in the kernel space. Second, in the kernel space, some examples have very similar conditional probabilities and can be used to construct transfer learning models. Third, the error rate of the transfer classifier can be bounded (Section 3). Thus, under a suitable kernel mapping space, two domains significantly different in their original space can both have similar marginal and conditional distributions, leading to effective knowledge transfer. Consider a synthetic example in Figure 1 where $\square/*$ denotes positive/negative. Figure 1(a) plots the target-domain data, "two circles", and the maximal margin decision boundary is the dashed ellipse. Figure 1(b) shows a source-domain data set, "two moons", where the decision boundary is the dashed curve. Obviously, two moons and two circles have significantly different distributions in the original space. However, after we map them into the kernel

KDD'09, June 28–July 1, 2009, Paris, France.