Information Sciences 298 (2015) 237–256

Contents lists available at ScienceDirect

Information Sciences

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

Sub-domain adaptation learning methodology

Jun Gao [a,b,c,*], Rong Huang [a], Hanxiong Li [c]

[a] School of Information Eng., Yancheng Institute of Technology, Yancheng, China
[b] School of Automation, Southeast University, Nanjing, China
[c] Department of System Eng. & Eng. Management, City University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 7 April 2013
Received in revised form 15 November 2014
Accepted 27 November 2014
Available online 5 December 2014

Keywords:
Maximum mean discrepancy
Local weighted mean
Projected maximum local weighted mean discrepancy
Multi-label classification
Support vector machines

ABSTRACT

Regarded as global methods, Maximum Mean Discrepancy (MMD) based transfer learning frameworks only reflect the global distribution discrepancy and structural differences between domains; they can reflect neither the inner local distribution discrepancy nor the structural differences between domains. To address this problem, a novel transfer learning framework with local learning ability, a Sub-domain Adaptation Learning Framework (SDAL), is proposed. In this framework, a Projected Maximum Local Weighted Mean Discrepancy (PMLMD) is constructed by integrating the theory and method of Local Weighted Mean (LWM) into MMD. PMLMD reflects global distribution discrepancy between domains through accumulating local distribution discrepancies between the local sub-domains in domains. In particular, we formulate in theory that PMLMD is one of the generalized measures of MMD. On the basis of SDAL, two novel methods are proposed by using Multi-label Classifiers (MLC) and Support Vector Machine (SVM). Finally, tests on artificial data sets, high dimensional text data sets and face data sets show the SDAL-based transfer learning methods are superior to or at least comparable with benchmarking methods.

1. Introduction

A major assumption in traditional statistical models is that the training data and the future data must have the same distribution; that is, the training data and the test data are Identically and Independently Distributed (I.I.D).[1] However, when the distribution appears non-identical, almost all the traditional intellectual learning models have to be rebuilt subject to the future data. In real-world applications, it is common that the data—such as cross-language texts, biological information, social internet information and multi-task studies [20]—may be non-I.I.D. The key challenge of these applications is that accurately labeled task-specific data are scarce while task- relevant data are abundant. Learning with non-I.I.D. data in such scenarios helps build accurate models by leveraging relevant data to perform new learning tasks, identifying the true connections among samples and their labels, and expediting the knowledge discovery process by simplifying the expensive data collection process. For such cases, Transfer Learning (TL) or Knowledge Transfer algorithms are proposed [23,29,4,8,27,12,19]. The aim of TL algorithms is to effectively build a statistical learning model that can deal with the target domain by using the knowledge obtained from the source domain [25]. These algorithms focus on knowledge transfer between different tasks or domains instead of simply

* Corresponding author at: School of Information Eng., Yancheng Institute of Technology, Yancheng, China.
E-mail address: gj0104211@163.com (J. Gao).

[1] In probability theory and statistics, a sequence or other collection of random variables is independent and identically distributed (I.I.D.) if each random variable has the same probability distribution as the others and all are mutually independent.

http://dx.doi.org/10.1016/j.ins.2014.11.041
0020-0255/

generalizing of cross-problems learning methods. From this perspective, TL algorithms differ greatly from traditional supervised or semi-supervised [34] or unsupervised methods since the latter deal with the training and the test data drawn from the same distribution but the former handle the source domains and the target ones having different distributions.

Domain Adaptation Learning (DAL), a special TL method, addresses how to build a statistical learning model after learning the knowledge from the source and target domains, which have different but related distributions. In DAL, the major calculating problem is how to minimize the distribution discrepancy between the task domains. In such cases, we need an effective measure that can reflect the distribution discrepancy between the task domains—this becomes the greatest challenge in building an effective DAL model. Recently, researchers have proposed some methods to judge the distribution discrepancy between the two domains, such as Kullback–Leibler Distance (KL-distance) [30] and Maximum Mean Discrepancy (MMD) [3]. KL-distance is an estimating method having parameters; it requires continual prior density estimates in the process of measuring the distribution discrepancy of the domains. But MMD is a measure having no parameters; it reflects the distribution discrepancy between the source and target domains by calculating their mean difference between them, thus making MMD simple, effective and intuitive. Based on MMD, some traditional methods—such as Transductive Support Vector Machine (TSVM) [15], Multi-label Classifiers (MLC) [14], and Feature Selection,—have been rebuilt to address a few domain adaptation learning problems. Furthermore, based on MMD, Quanz and Huan [24] proposed and utilized Projected Maximum Mean Discrepancy (PMMD)[2] to show the distribution discrepancy between the embedded domains of the source and target domains.

Either MMD or PMMD shows the distribution discrepancy between different domains in the form of the population mean difference between the domains or between the corresponding embedded subspaces of the domains. As stated in statistical theories [33], the population mean or the expectation of a domain, as an effective statistical feature, often indicates the global distribution and structure information of the domain. And from the perspective of geometry, the population mean shows the distribution of spatial data better; that is, it can better reflect the statistical feature of Gaussian distribution data. So MMD and PMMD reflect to some extent the discrepancy of the global distribution or the global structure information between the source domain and the target domain. Then they are more suitable for reflecting the distribution discrepancy between domains having an apparent Gaussian distribution. On this level, the MMD-based domain adaptation learning methods, such as Multi-view Transfer Learning with a Large Margin Framework (MVTL-LM) [37], Domain Transfer Multiple Kernel Learning Framework (DTMKL) [9,10], Domain Adaptation Support Vector Machine (DASVM) [5], Domain Adaptation Kerneled Support Vector Machine (DAKSVM) [28], Maximum Mean Discrepancy Embedding (MMDE) [21], and Multi-label Classification Learning Framework (MCLF) [6], are in the category of global methods; hence, to some extent, they ignore the local discrepancy between domains and the local structural information of different domains. So far, little research has been reported to address this problem through using the novel measure with local learning ability.

Therefore, in this paper, we propose a novel transfer learning framework with local learning ability: Sub-domain Adaptation Learning Framework (SDAL). In this framework, we integrate Local Weighted Mean (LWM) [2,35] into MMD and then propose a novel criterion that can effectively measure the distribution discrepancy between the source domain and the target domain—Projected Maximum Local Weighted Mean Discrepancy (PMLMD). Finally, in the SDAL framework, by using MLC and SVM, two sub-domain adaptation learning methods are constructed for TL problems. The framework in this paper has the following advantages:

(1) PMLMD based on LWM and MMD cannot only effectively calculate the local distribution discrepancy between sub-domains in different domains but also effectively reflect the local geometrical discrepancy in different domains. Furthermore, it can reflect the global distribution and geometrical discrepancy between domains through accumulating the local distribution discrepancies between sub-domains. We theoretically formulate that PMLMD is a generalization of MMD and PMMD. Additionally, a novel definition, Closest Local Sub-domain (CLSD), is presented to justify how to calculate the local distribution discrepancy between the source and target domains.
(2) SDAL welcomes many traditional statistical learning methods, such as SVM, Support Vector Regression (SVR) [38], and MLC, to solve the problems of domain adaptation learning. In particular, in the SDAL framework, we use MLC and SVM to build two sub-domain adaptation learning methods for TL. The former, MLC based sub-domain adaptation learning (MLC-SDAL), is a local linear multi-label classification sub-domain adaptation learning method, which is an effective classifier and can realize low-dimensional embedding of the original input space corresponding to the source and target domains, so it can be taken as the generalized form of MLC and MCLF. The latter, SVM- based sub-domain learning (SVM-SDAL), can be not only a large-margin domain support vector method but also a local learning classifier; on this level, it inherits and extends all the advantages of TSVM and above MMD-based support vector machine algorithms. SDAL of course can address the traditional supervised, semi-supervised or unsupervised intellectual recognition problems.
(3) All the advantages of SDAL are justified by the tests on artificial data with obvious local manifold feature, high-dimension text data and face recognition data.

[2] PMMD and MMD differ from each other in that PMMD reflects the distribution discrepancy between the low dimensional spaces embedded in the source domain and the target domain respectively while MMD reflects the distribution discrepancy between the source domain and the target domain.