Journal of Fiber Bioengineering and Informatics 8:1 (2015) 105–115
doi:10.3993/jfbi03201510

Ensemble Inductive Transfer Learning ⋆

Xiaobo Liu [a,*], Guangjun Wang [a], Zhihua Cai [b], Harry Zhang [c]

[a] *School of Automation, China University of Geosciences, Wuhan 430074, China*

[b] *School of Computer Science, China University of Geosciences, Wuhan 430074, China*

[c] *Faculty of Computer Science, University of New Brunswick, Fredericton, NB, E3B5A3, Canada*

Abstract

Inductive transfer learning is a major research area in transfer learning which aims at achieving a high performance in the target domain by inducing the useful knowledge from the source domain. By combining decisions from individual classifiers, ensemble learning can usually reduce variance and achieve higher accuracy than a single classifier. In this paper, we propose a novel Ensemble Inductive Transfer Learning (EITL) method. EITL builds a set of classifiers by recording the iterative process of knowledge transfer. In each iteration, it uses the classifier of the source domain, the base classifier of the target domain built on the initial labeled data, and the most recent classifier built on the updated labeled data, to classify unlabeled instances, and add some self-labeled instances to the labeled data, and then trains a new classifier. At the end, all the classifiers built in this process are used for classification. We conduct experiments on synthetic data sets and six UCI data sets, which show that EITL is an effective algorithm in terms of classification accuracy.

Keywords: Transfer Learning; Ensemble Learning; Machine Learning

1 Introduction

In the machine learning field, a major challenge is that labeled data is easy to outdate, and data labeling is often expensive and time-consuming. One direct consequence is the lack of training data, which may result in an unsatisfactory performance by using the traditional machine learning methods to construct a model. Moreover, the feature space or distributions may change over time. Transfer learning is a new research area in machine learning, which provides a new approach to tackle this issue. Transfer learning reuses the useful certain parts of auxiliary data sets to train a classifier for the new data, although the auxiliary data sets may have different feature spaces or different distributions [12]. We called the outdated data sets or the auxiliary data sets as the source domain data sets. Usually, there are a variety of approaches to transfer knowledge from source domain to target domain, such as transferring some useful instances [3], feature

⋆Project supported by the Key Project of the Natural Science Foundation of Hubei Province, China (No. 2013CFA004) and the National Natural Science Foundation of China (No. 61403351).

*Corresponding author.
Email address: xbliu@cug.edu.cn (Xiaobo Liu).

representation [13], parameters [2], or relational knowledge [11]. In our method, we utilize the source domain to help the target domain to label the unlabeled data in the target domain.

Ensemble approach is usually more reliable than single approach to make classification [17]. Our motivation is that, we construct a set of classifiers and then classify unlabeled data by taking a vote on their predictions [4]. The initial ensemble includes two distinct classifiers built on the source domain and the target domain with a few initial labeled data, which are used for partly classifying the unlabeled data in target domain. The resulting labeled data is combined with the initial labeled data in the target domain to create a third classifier of the ensemble. The third classifier is updated with newly labeled data at each successive iteration. After many iterations, we have generated a set of classifiers for the target domain. Then these classifiers are used to predict the labels of test data by the majority vote strategy [8].

The main contributions of this paper are: (1) We use the ensemble method to repeatedly label the unlabeled data in the target domain, which can improve the performance of the target domain's task – classification. Meanwhile, the source domain takes part in the decision at each iteration, which can always play a role in the ensemble. (2) Our method is useful to do transfer learning. When only a few labeled data is given in the target domain, our method successfully use the source domain to help the target domain to label the unlabeled data in the target domain. Thus, the smaller the number of labeled data in the target domain is, the more useful our method is.

The rest of this paper is organized as follows: Section 2 presents related work about transfer learning and ensemble machine learning. The detailed introduction of our improved algorithm is described in Section 3. Section 4 displays the experiments on synthetic data sets and six UCI data sets, and analyzes the results. And Section 5 gives the conclusions of our work.

2 Related Work

The source domain data usually has a different distribution from the target domain data. If we reuse the source domain data directly, it will be unfeasible. If we discard the source domain data completely, the few labeled instances in the target domain are insufficient to train a good classifier. How to deal with those situations is the major challenge in transfer learning.

Dai et al. [3] proposed the TrAdaBoost method which iteratively re-weights the source domain data to reduce the effect of the *bad* source data while encouraging the *good* source data to contribute more to the target domain. Shi et al. [14] proposed a framework to actively transfer the knowledge from the source domain to help learn the target domain, and query experts only when necessary. Jiang et al. [6] proposed a general instance weighting framework for domain adaptation, which removed *misleading* source domain instances and added labeled target domain instances with higher weights. Liao et al. [9] proposed *Migratory-Logit* algorithm which is a new active learning approach for selecting the labeled examples in a target domain.

Constructing a good ensemble of classifiers has been an active research area in machine learning [4]. By combining decisions from individual classifiers, ensembles can usually reduce variance and achieve higher accuracy than an individual classifier.

Kamishima et al. [7] proposed a TrBagging method which is an extension of bagging. In order to reuse certain parts of the data in the source domain to benefit the learning in the target domain, the authors combined the source domain data sets and the target domain data sets, used