Pattern Recognition Letters 88 (2017) 64–71

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Testing exchangeability for transfer decision

Shuang Zhou*, Evgueni Smirnov, Gijs Schoenmakers, Kurt Driessens, Ralf Peeters

Department of Data Science and Knowledge Engineering, Maastricht University, P.O.BOX 616, Maastricht, 6200 MD, The Netherlands

ARTICLE INFO

Article history:
Received 23 March 2016
Available online 16 January 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Instance-transfer learning
Conformity prediction framework
Exchangeability test

ABSTRACT

This paper introduces a non-parametric test to decide whether to transfer data from a source domain to a target domain to improve the generalization performance of predictive models on the target domain. The test is based on the conformal prediction framework: it statistically tests whether the target and source data are generated from the same distribution under the exchangeability assumption. The experiments show that the test is capable of outperforming existing methods when it decides on instance transfer.

1. Introduction

Instance transfer has received significant attention in the last decade [19]. The goal is to improve the predictive models for a *target* domain by exploiting data from a (closely) related *source* domain. A thorough analysis of instance transfer [22] shows that its effectiveness depends on the relevance of the source domain to the target domain. As a result, a critical problem we have in practice is to decide whether we can transfer the source data while training predictive models for the target one. Adequately addressing this problem guarantees that we avoid negative transfer when adopting source data degrades the performance of the final models [19] .

The standard approach to the problem of deciding whether to transfer source data has been proposed in a number of works all based on the same principle [6,7,21,26]. This approach considers the distance between the target and source probability distributions as the difference between the target and source domains. Thus, the source data is transferred iff the distance between the probability distributions is small enough. There are however two major drawbacks of this approach. First, it is sensitive to the accuracy of estimating target and source probability distributions, since the real distributions are usually unknown. When this accuracy degrades (for example when the data is limited in size or high dimensional), the approach can be misleading. Second, the bound used to select or disregard source data is set by the user in an ad-hoc fashion and the approach does not lend itself to making statistically sound decisions; e.g., to transfer only if the target and source probability distributions are similar with high probability. This is because the distance values provided by the approach are very difficult to relate to the null hypothesis that “the target and source probability distributions are similar” in a statistical sense. This can easily result in negative transfer.

In this paper, we propose to avoid the aforementioned drawbacks by using a non-parametric test to decide on transfer using source domain based on the conformal prediction framework [20]. It tests whether the target data and the source data have been generated from the same distribution under the exchangeability assumption [2] and makes it possible to decide on transfer using an interpretable significance level.

The essential part of the new test is a p-value function that can be used to estimate the relevance of the target and source data in a statistically sound way. For any target and source data the function returns a p-value related to the null hypothesis “the target and source data have been generated from the target distribution under the exchangeability assumption”. The function can be instantiated dependent/independent on/of the predictive models used. The validity of the function is proven.

The rest of the paper is organized as follows. Section 2 formalizes the classification task and instance transfer task, and provides an overview of related work. Our conformity-based test to decide on transfer from source data is introduced in Section 3. Section 4 provides an experimental comparison with existing approaches. Finally, Section 5 concludes the paper.

* Corresponding author.
E-mail address: shuang.zhou@maastrichtuniversity.nl (S. Zhou).

http://dx.doi.org/10.1016/j.patrec.2016.12.021
0167-8655/

2. Classification and instance transfer

In this section we present classification tasks in the context of instance transfer learning and discuss past work.

2.1. Task definition

Let X be a feature space and Y be a class set. A domain is defined as a 2-tuple consisting of a labeled space $(X \times Y)$ and a probability distribution P_{XY} over the labeled space[1]. We consider first a domain $\langle (X \times Y), P^t_{XY} \rangle$ that we call a target domain. The target data set T is a finite multi set of m_T instances $(x_t, y_t) \in X \times Y$ drawn from the target distribution P^t_{XY} under the randomness assumption. Given a test instance $x_{m_T+1} \in X$, the target classification task is to find an estimate $\hat{y}_{m_T+1} \in Y$ for the true class of x_{m_T+1} according to P^t_{XY}.

Let us consider a second domain $\langle (X \times Y), P^s_{XY} \rangle$ that we call a source domain. The source data S is a finite multi set of m_S instances $(x_s, y_s) \in X \times Y$ drawn from the source distribution P^s_{XY} under the randomness assumption. Knowing that the target domain and the source domain are relevant, we define *the instance-transfer classification task* as a classification task with an auxiliary source data set S in addition to the target data set T. We note that the class of a new test instance is estimated according to the target distribution P^t_{XY}.

From the definition above it follows that instance transfer is sensitive to the relevance of the source data S to the target data T. Hence, the problem of deciding whether we can transfer the source data in order to improve class estimation in the target domain is important for its overall success.

2.2. Related work

As it is mentioned in Section 1, the standard approach of deciding whether to transfer source data consists of two stages. First, the target and source probability distributions are estimated, and then the distance between the distributions is computed. Based on the distance, the transfer decision is made. Below the main methods within this approach are described.

By assuming that the underlying distributions are normal, the distance can be estimated using the Mahalanobis distance, which is a common statistical distance between an instance and a distribution [8]. It measures how many standard deviations the instance is away from the mean of the distribution. Given target and source data, the distance between these two sets can be estimated by averaging the Mahalanobis distance of each source instance to the target distribution (estimated from the target data). Although the measure is widely used, it assumes a normal distribution of the target instances. In practice, this assumption does not often hold.

Kullback–Leibler divergence (KL-divergence) is one of the most widely used measures to compare probability distributions [14]. Given target and source data, KL-divergence from the P^s_{XY} to P^t_{XY} is defined as the expectation of the logarithmic quotient of the estimated target and source densities, where the expectation is taken w.r.t. the target density. Some applications of KL-divergence can be found in [1,6,7,27]. In [6], by assuming the independence between features, the KL-divergence from P^s_{XY} to P^t_{XY} was computed as the sum of feature-level KL-divergences. In [27], P^s_{XY} and P^t_{XY} were assumed to be mixtures of Gaussians. The KL-divergence was then calculated based on two Gaussian mixtures estimated from T and S. KL-divergence assumes that the probability densities can be estimated precisely from the data. When the number of the measurements is small and/or the data-space is highly dimensional the approximations can result in an inaccurate KL-divergence estimation.

The $\mathcal{A}$-distance was introduced by [12]. Given target and source probability distributions P^t_{XY} and P^s_{XY}, and a collection $\mathcal{A}$ of data sets, the $\mathcal{A}$-distance between P^t_{XY} and P^s_{XY} is defined as the upper bound of the absolute difference of the probabilities of generating sets $A \in \mathcal{A}$ w.r.t. P^t_{XY} and P^s_{XY}. The $\mathcal{A}$-distance depends on the choice of the sets collection $\mathcal{A}$, and determining a good collection is an open problem.

The discrepancy distance, proposed by Mansour et al. [17], estimates the difference between the target and source conditional distributions $P^t_{Y|X}$ and $P^s_{Y|X}$ from the perspective of a hypothesis space H. The key idea is that the target (source) classifier $h^t \in H$ ($h^s \in H$) based on the target (source) data sets T (S) can be used to approximate the conditional target (source) distribution $P^t_{Y|X}$ ($P^s_{Y|X}$). Therefore, the discrepancy distance is computed as the disagreement between the target and source classifier h^t and h^s by labeling instances from the union of target and source data. One drawback of using the discrepancy distance is that the difference between the target and source domains is estimated only in terms of the difference in conditional distributions without taking into account the difference between the marginal distributions.

Since any joint distribution P_{XY} can be given as the product of marginal distribution P_X and the conditional distribution $P_{Y|X}$, the transfer cross validation framework (TrCV) [26] measures the distance between marginal distributions and conditional distributions and then combines them to indicate the joint distribution discrepancy. More precisely, applying TrCV is a two-step process. First, a density ratio weighting approach is used to assess the difference in marginal distributions P^t_X and P^s_X. Second, a reverse validation framework is employed to quantify the discrepancy between conditional probabilities $P^t_{Y|X}$ and $P^s_{Y|X}$. The distance between target and source joint distributions is then calculated as the product of marginal discrepancy and conditional discrepancy.

As it is mentioned in Section 1, all these methods are sensitive to the accuracy of distribution estimation, and do not support instance transfer decision in a statistical sense. To avoid these problems, in the next section we propose our solution.

3. A conformity-based test for transfer decisions

We propose a non-parametric statistical test to decide on instance transfer from given source data. In the original problem formulation, target and source data are generated under the randomness assumption. This assumption leads to a null hypothesis that the joint data set $T \cup S$ was generated from the target probability distribution P^t_{XY} under the randomness assumption. Thus, one could employ some of the randomness tests from the algorithmic theory of randomness [5,18,25]. However, it is a well-known fact that those tests are incomputable [24].

To go around the computability problem of the randomness tests we propose to employ the conformal prediction framework [20] instead. In this context we introduce a conformity-based test to decide on transfer from the source data. The key idea stays the same but under the exchangeability assumption of data generation [2] that treats the data-sets as finite sequences sampled from the probability distributions. The null hypothesis then becomes that the data sequence TS consisting first of target data sequence T and followed by source data sequence S was generated from the target probability distribution P^t_{XY} under the exchangeability assumption. If the null hypothesis is accepted at some significance level, it implies that at that level the target and source data sequences T and S are relevant, and the source data sequence S can be transferred. Otherwise, the target data sequence T should be used on its own. This way we avoid probability-distribution estimations and provide

[1] For the sake of completeness the marginal distribution over X is denoted by P_X, and the conditional distribution over Y given X by $P_{Y|X}$.