Optimal Transport for Domain Adaptation

Nicolas Courty, Rémi Flamary, Devis Tuia, *Senior Member, IEEE*,
Alain Rakotomamonjy, *Member, IEEE*

Abstract—Domain adaptation is one of the most challenging tasks of modern data analytics. If the adaptation is done correctly, models built on a specific data representation become more robust when confronted to data depicting the same classes, but described by another observation system. Among the many strategies proposed, finding domain-invariant representations has shown excellent properties, in particular since it allows to train a unique classifier effective in all domains. In this paper, we propose a regularized unsupervised optimal transportation model to perform the alignment of the representations in the source and target domains. We learn a transportation plan matching both PDFs, which constrains labeled samples of the same class in the source domain to remain close during transport. This way, we exploit at the same time the labeled samples in the source and the distributions observed in both domains. Experiments on toy and challenging real visual adaptation examples show the interest of the method, that consistently outperforms state of the art approaches. In addition, numerical experiments show that our approach leads to better performances on domain invariant deep learning features and can be easily adapted to the semi-supervised case where few labeled samples are available in the target domain.

Index Terms—Unsupervised Domain Adaptation, Optimal Transport, Transfer Learning, Visual Adaptation, Classification.

1 Introduction

Modern data analytics are based on the availability of large volumes of data, sensed by a variety of acquisition devices and at high temporal frequency. But this large amounts of heterogeneous data also make the task of learning semantic concepts more difficult, since the data used for learning a decision function and those used for inference tend not to follow the same distribution. Discrepancies (also known as drift) in data distribution are due to several reasons and are application-dependent. In computer vision, this problem is known as the visual adaptation domain problem, where domain drifts occur when changing lighting conditions, acquisition devices, or by considering the presence or absence of backgrounds. In speech processing, learning from one speaker and trying to deploy an application targeted to a wide public may also be hindered by the differences in background noise, tone or gender of the speaker. In remote sensing image analysis, one would like to leverage from labels defined over one city image to classify the land occupation of another city. The drifts observed in the probability density function (PDF) of remote sensing images are caused by variety of factors: different corrections for atmospheric scattering, daylight conditions at the hour of acquisition or even slight changes in the chemical composition of the materials.

For those reasons, several works have coped with these drift problems by developing learning methods able to transfer knowledge from a source domain to a target domain for which data have different PDFs. Learning in this PDF discrepancy context is denoted as the domain adaptation problem [37]. In this work, we address the most difficult variant of this problem, denoted as **unsupervised domain adaptation**, where data labels are only available in the source domain. We tackle this problem by assuming that the effects of the drifts can be reduced if data undergo a phase of *adaptation* (typically, a non-linear mapping) where both domains look more alike.

Several theoretical works [2], [36], [22] have emphasized the role played by the divergence between the data probability distribution functions of the domains. These works have led to a principled way of solving the domain adaptation problem: transform data so as to make their distributions "closer", and use the label information available in the source domain to learn a classifier in the transformed domain, which can be applied to the target domain. Our work follows the same intuition and proposes a transformation of the source data that fits **a least effort principle**, *i.e.* an effect that is minimal with respect to a transformation cost or metric. In this sense, the adaptation problem boils down to: *i)* finding a transformation of the input data matching the source and target distributions and then *ii)* learning a new classifier from the transformed source samples. This process is depicted in Figure 1. In this paper, we advocate a solution for finding this transformation based on *optimal transport*.

Optimal Transport (OT) problems have recently raised interest in several fields, in particular because OT theory can be used for computing distances between probability distributions. Those distances, known under several names in the literature (Wasserstein, Monge-Kantorovich or Earth Mover distances) have important properties: *i)* They can be evaluated directly on empirical estimates of the distribu-

Manuscript received January 2015; revised September 2015.

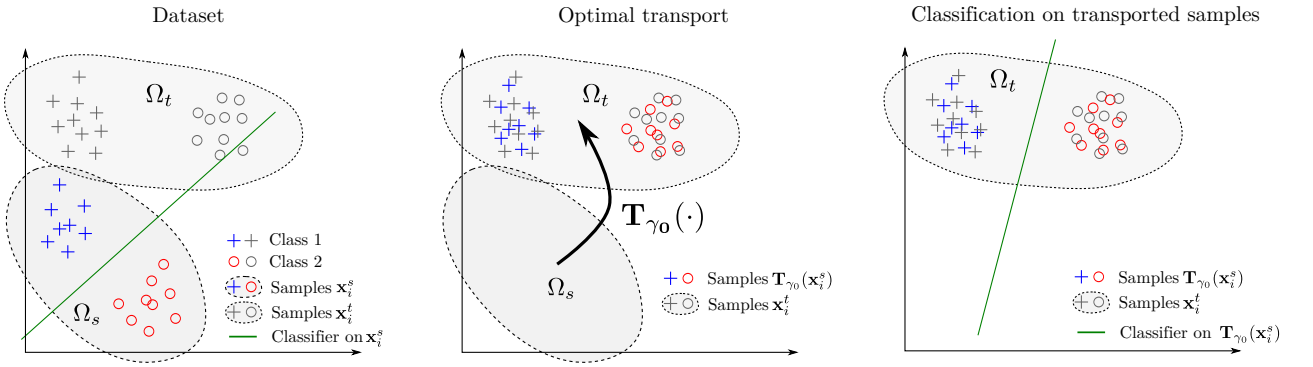

Fig. 1: Illustration of the proposed approach for domain adaptation. (left) dataset for training, *i.e.* source domain, and testing, *i.e.* target domain. Note that a classifier estimated on the training examples clearly does not fit the target data. (middle) a data dependent transportation map $\mathbf{T}_{\gamma 0}$ is estimated and used to transport the training samples onto the target domain. Note that this transformation is usually not linear. (right) the transported labeled samples are used for estimating a classifier in the target domain.

tions without having to smoothen them using non-parametric or semi-parametric approaches; *ii)* By exploiting the geometry of the underlying metric space, they provide meaningful distances even when the supports of the distributions do not overlap. Leveraging from these properties, we introduce a novel framework for unsupervised domain adaptation, which consists in learning an optimal transportation based on empirical observations. In addition, we propose several regularization terms that favor learning of better transformations *w.r.t.* the adaptation problem. They can either encode class information contained in the source domain or promote the preservation of neighborhood structures. An efficient algorithm is proposed for solving the resulting regularized optimal transport optimization problem. Finally, this framework can also easily be extended to the semi-supervised case, where few labels are available in the target domain, by a simple and elegant modification in the optimal transport optimization problem.

The remainder of this Section presents related works, while Section 2 formalizes the problem of unsupervised domain adaptation and discusses the use of optimal transport for its resolution. Section 3 introduces optimal transport and its regularized version. Section 4 presents the proposed regularization terms tailored to fit the domain adaptation constraints. Section 5 discusses algorithms for solving the regularized optimal transport problem efficiently. Section 6 evaluates the relevance of our domain adaptation framework through both synthetic and real-world examples.

1.1 Related works

Domain adaptation. Domain adaptation strategies can be roughly divided in two families, depending on whether they assume the presence of few labels in the target domain (semi-supervised DA) or not (unsupervised DA).

In the first family, methods which have been proposed include searching for projections that are discriminative in both domains by using inner products between source samples and transformed target samples [42], [32], [29]. Learning projections, for which labeled samples of the target domain fall on the correct side of a large margin classifier trained on the source data, have also been proposed [27]. Several works based on extraction of common features under pairwise constraints have also been introduced as domain adaptation strategies [26], [52], [47].

The second family tackles the domain adaptation problem assuming, as in this paper, that no labels are available in the target domain. Besides works dealing with sample reweighting [46], many works have considered finding a common feature representation for the two (or more) domains. Since the representation, or *latent space*, is common to all domains, projected labeled samples from the source domain can be used to train a classifier that is general [18], [38]. A common strategy is to propose methods that aim at finding representations in which domains match in some sense. For instance, adaptation can be performed by matching the means of the domains in the feature space [38], aligning the domains by their correlations [33] or by using pairwise constraints [51]. In most of these works, feature extraction is the key tool for finding a common latent space that embeds discriminative information shared by all domains.

Recently, the unsupervised domain adaptation problem has been revisited by considering strategies based on a gradual alignment of a feature representation. In [24], authors start from the hypothesis that domain adaptation can be better estimated when comparing gradual distortions. Therefore, they use intermediary projections of both domains along the Grassmannian geodesic connecting the source and target eigenvectors. In [23], [54], all sets of transformed intermediary domains are obtained by using

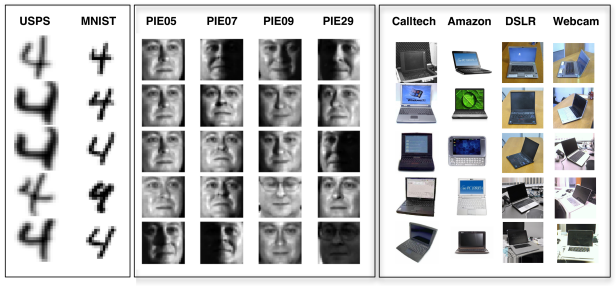

Fig. 4: Examples from the datasets used in the visual adaptation experiment. 5 random samples from one class are given for all the considered domains.

- **PCA**, which consists in applying a projection on the first principal components of the joint source/target distribution (estimated from the concatenation of source and target samples);
- **GFK**, Geodesic Flow Kernel [23];
- **TSL**, Transfer Subspace Learning [44], which operates by minimizing the Bregman divergence between the domains embedded in lower dimensional spaces;
- **JDA**, Joint Distribution Adaptation [34], which extends the Transfer Component Analysis algorithm [38];

In unsupervised DA no target labels are available. As a consequence, it is impossible to consider a cross-validation step for the hyper-parameters of the different methods. However, and in order to compare the methods fairly, we follow the following protocol. For each source domain, a random selection of 20 samples per class (with the only exception of 8 for the DSLR dataset) is adopted. Then the target domain is equivalently partitioned in a validation and test sets. The validation set is used to obtain the best accuracy in the range of the possible hyper-parameters. The accuracy, measured as the percent of correct classification over all the classes, is then evaluated on the testing set, with the best selected hyper-parameters. This strategy normally prevents overfitting on the testing set. The experimentation is conducted 10 times, and the mean accuracy over all these realizations is reported.

We considered the following parameter range : for subspace learning methods (**PCA**,**TSL**, **GFK**, and **JDA**) we considered reduced k-dimensional spaces with $k \in \{10, 20, \dots, 70\}$. A linear kernel was chosen for all the methods with a kernel formulation. For the all methods requiring a regularization parameter, the best value was searched in $\lambda = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. The λ and η parameters of our different regularizers (Equation (16)), are validated using the same search interval. In the case of the Laplacian regularization (**OT-Laplace**), $\mathbf{S}_t$ is a binary matrix which encodes a nearest neighbors graph with a 8-connectivity. For the source domain, $\mathbf{S}_s$ is filtered such that connections between elements of different classes are pruned. Finally, we set the α value Equation (20) to 0.5.

6.2.3 Results on unsupervised domain adaptation

Results of the experiment are reported in Table 3 where the best performing method for each domain adaptation problem is highlighted in bold. On average, all the OT-based domain adaptation methods perform better than the baseline methods, except in the case of the PIE dataset, where **JDA** outperforms the OT-based methods in 7 out of 12 domain pairs. A possible explanation is that the dataset contains a lot of classes (68), and the EM-like step of **JDA**, which allows to take into account the current results of classification on the target, is clearly leading to a benefit. We notice that **TSL**, which is based on a similar principle of distribution divergence minimization, almost never outperforms our regularized strategies, except on pair A→C. Among the different optimal transport strategies, **OT-Exact** leads to the lowest performances. **OT-IT**, the entropy regularized version of the transport, is substantially better than **OT-Exact**, but is still inferior to the class-based regularized strategies proposed in this paper. The best performing strategies are clearly **OT-GL** and **OT-Laplace** with a slight advantage for **OT-GL**. **OT-LpL1**, which is based on a similar regularization strategy as **OT-GL**, but with a different optimization scheme, has globally inferior performances, except on some pairs of domains (e.g. C→A) where it achieves better scores. On both digits and objects recognition tasks, **OT-GL** significantly outperforms the baseline methods.

In the next experiment (Table 4), we use the same experimental protocol on different features produced by the DeCAF deep learning architecture [19]. We report the results of the experiment conducted on the Office-Caltech dataset, with the **OT-IT** and **OT-GL** regularization strategies. For comparison purposes, **JDA** is also considered for this adaptation task. The results show that, even though the deep learning features yield naturally a strong improvement over the classical SURF features, the proposed OT methods are still capable of improving significantly the performances of the final classification (up to more than 20 points in some case, e.g. D→A or A→W). This clearly shows how OT has the capacity to handle non-stationarity in the distributions that the deep architecture has difficulty handling. We also note that using the features from the 7th layer instead of the 6th does not bring a strong improvement in the classification accuracy, suggesting that part of the work of the 7th layer is already performed by the optimal transport.

6.2.4 Semi-supervised domain adaptation

In this last experiment, we assume that few labels are available in the target domain. We thus benchmark