

On the Usefulness of Similarity Based Projection Spaces for Transfer Learning

Emilie Morvant, Amaury Habrard, Stéphane Ayache

▶ **To cite this version:**

Emilie Morvant, Amaury Habrard, Stéphane Ayache. On the Usefulness of Similarity Based Projection Spaces for Transfer Learning. First International Workshop on Similarity-Based Pattern Recognition, Sep 2011, Venise, Italy. pp.1-16. hal-00628991

HAL Id: hal-00628991
https://hal.science/hal-00628991

Submitted on 4 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the Usefulness of Similarity based Projection Spaces for Transfer Learning*

Emilie Morvant, Amaury Habrard, and Stéphane Ayache

Laboratoire d'Informatique Fondamentale de Marseille, Aix-Marseille Université,
CNRS UMR 6166, 13453 Marseille cedex 13, France
{emilie.morvant,amaury.habrard,stephane.ayache}@lif.univ-mrs.fr

Abstract. Similarity functions are widely used in many machine learning or pattern recognition tasks. We consider here a recent framework for binary classification, proposed by Balcan et al., allowing to learn in a potentially non geometrical space based on good similarity functions. This framework is a generalization of the notion of kernels used in support vector machines in the sense that allows one to use similarity functions that do not need to be positive semi-definite nor symmetric. The similarities are then used to define an explicit projection space where a linear classifier with good generalization properties can be learned. In this paper, we propose to study experimentally the usefulness of similarity based projection spaces for transfer learning issues. More precisely, we consider the problem of domain adaptation where the distributions generating learning data and test data are somewhat different. We stand in the case where no information on the test labels is available. We show that a simple renormalization of a good similarity function taking into account the test data allows us to learn classifiers more performing on the target distribution for difficult adaptation problems. Moreover, this normalization always helps to improve the model when we try to regularize the similarity based projection space in order to move closer the two distributions. We provide experiments on a toy problem and on a real image annotation task.

Keywords: Good Similarity Functions, Transfer Learning, Domain Adaptation, Image Classification

1 Introduction

Many machine learning or pattern recognition algorithms are based on similarity functions. Among all of the existing methods, we can cite the famous k-nearest neighbors, k-means or support vector machines (SVM). An important point is to choose or adapt the similarity to the problem considered. For example, approaches dealing with numerical vectors are often based on the Mahalanobis

* This work was supported in part by the french project VideoSense ANR-09-CORD-026 of the ANR in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

distance [12, 15, 27] and many methods designed for structured data (strings, trees or graphs) exploit the notion of edit distance [7, 14, 24]. For binary classification with SVM classifiers, the similarity function must often be a valid kernel[1] in order to define a potentially implicit projection space which is an Hilbert space and where data can be more easily separated. In this case, the similarity function must be symmetric and positive semi-definite (PSD), allowing one to define a valid dot product in the implicit projection space. However, these constraints may rule out some natural similarity functions. Recently, a framework proposed by Balcan *et al.* [2, 3] considers a notion of *good similarity function* that overcomes these limitations. Intuitively, this notion only requires that a sufficient amount of examples are on average more similar to a set of *reasonable* points of the same class than to *reasonable* points of the opposite class. Then, the similarity can be used to build an explicit (potentially non geometrical) projection space, corresponding to the vector of similarities to the reasonable examples. In this similarity based projection space, a classifier with good generalization capabilities can be learned.

This kind of result holds in a classical machine learning setting, where test data are supposed to have been generated according to the same distribution than the one used for generating labeled learning data. This assumption is in fact very useful to obtain good generalization results, but is not always valid in every application. For example, in an image classification task, if labeled data consist of images extracted from the web and test data images extracted from different videos, the various methods of data acquisition may imply that labeled data are no longer representative of test data and thus of the underlying classification task. This kind of issue is a special case of transfer learning [22] called *domain adaptation* (DA) [18, 23]. DA arises when learning and test data are generated according to two different probability distributions: the first one generating learning data is often referred to as the *source domain*, while the second one for test data corresponds to the *target domain*. According to the existing theoretical frameworks of DA [4, 20] a classifier can perform well on the target domain if its error relatively to the source distribution and the divergence between the source and target distributions are together low. One possible solution to learn a performing classifier on the target domain is to find a projection space in which the source and target distributions are close while keeping a low error on the source domain. Many approaches have been proposed in the literature to tackle this problem [9–11, 19].

In this paper, we consider the case where a learning algorithm is provided with labeled data from the source domain and unlabeled data from the target one. Our aim is to investigate the interest of the framework of Balcan *et al.* for domain adaptation problems. More precisely, we propose to study how we can use the lack of geometrical space of this framework to facilitate the adaptation. We consider two aspects. First, the influence of a renormalization of the similarity function according to the unlabeled source and target data. Second, the addition of a regularization term to the optimization problem considered for

[1] Nevertheless there exists some approaches allowing to use indefinite kernels [16].

learning the classifier in order to select reasonable points that are relevant for the adaptation. This approach can be seen as a feature selection for transfer learning aiming at moving closer the two distributions. We show experimentally that these two aspects can help to learn a better classifier for the target domain. Our experiments are based on a synthetic toy problem and on a real image annotation task.

The paper is organized as follows. We introduce some notations in Section 2. Then we present the framework of *good similarity functions* of Balcan *et al.* in Section 3. We next give a brief overview of *domain adaptation* in Section 4. We present in Section 5 the approach considered and we describe our experimental study in Section 6. We conclude in Section 7.

2 Notations

We denote by $X \subseteq \mathbb{R}^d$ the input space. We consider binary classification problems with $Y = \{-1, 1\}$, the label set. A learning task is modeled as a probability distribution P over $X \times Y$, D being the marginal distribution over X. For any labeled sample S drawn from P, we denote by $S_{|X}$ the sample constituted of all the instances of S without the labels. In a classical machine learning setting, the objective is then to learn a classifier $h : X \to Y$ belonging to a class of hypothesis $\mathcal{H}$ such that h has a low generalization error $\text{err}_P(h)$ over the distribution P. The generalization error $\text{err}_P(h)$ corresponds to the probability that h can commit an error according to the distribution P, which is defined as follows:

$$\forall h \in \mathcal{H},\ \text{err}_P(h) = \mathbb{E}_{(\mathbf{x},y)\sim P} L(h(\mathbf{x}), y)$$

where L corresponds to the loss function modeling the fact that $h(\mathbf{x}) \neq y$. We will see later that in a DA scenario, we consider two probability distributions P_S and P_T corresponding respectively to a source domain and a target one.

We now give a definition about the notion of similarity functions.

Definition 1. *A similarity function over X is any pairwise function*

$$K : X \times X \to [-1, 1].$$

K is symmetric if for any $\mathbf{x}, \mathbf{x}' \in X$: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$.

A similarity function is a valid kernel function if it is positive semi-definite, meaning that there exists a function ϕ from X to an implicit Hilbert space such that K defines a valid dot product in this space, *i.e.* $K(x, x') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$. Using a valid kernel offers the possibility to learn a good classifier into a high dimensional space where the data are supposed to be linearly separable. However, the choice or the definition of a good kernel can be a tricky task in general. We present in the next section a framework that proposes a rather intuitive notion of good similarity function that gets rid of the constraints of a kernel.

3 Learning with Good Similarity Functions

In this section, we present the class $\mathcal{H}$ of linear classifiers considered in this paper. These classifiers are based on a notion of *good similarity function* for a given classification task. A common general idea is that such a similarity function is able to separate examples of the same class from examples of the opposite class with a given confidence $\gamma > 0$. Given two labeled examples $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ of $X \times Y$, this idea can be formalized as follows: if $y = y'$ then $K(\mathbf{x}, \mathbf{x}') > \gamma$, otherwise we want $K(\mathbf{x}, \mathbf{x}') < -\gamma$. This can be summarized by the following formulation: $yy'K(\mathbf{x}, \mathbf{x}') > \gamma$. The recent learning framework proposed by Balcan *et al.* [2, 3], has generalized this idea by requiring the similarity to be good over a set of *reasonable* points.

Definition 2 (Balcan et al. [2]). *A similarity function K is an (ϵ,γ,τ)-*good similarity function *for a learning problem P if there exists a (random) indicator function $R(\mathbf{x})$ defining a set of reasonable points such that the following conditions hold:*

(i) A $1-\epsilon$ probability mass of examples $(\mathbf{x}, y)$ satisfy

$$\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[yy'K(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}')\big] \;\geq\; \gamma, \tag{1}$$

(ii) $Pr_{\mathbf{x}'}[R(\mathbf{x}')] \geq \tau$.

From this definition, a large proportion of examples must be on average more similar, with respect to the margin γ, to random reasonable examples of the same class than to random reasonable examples of the opposite class. Moreover, at least a proportion τ of examples should be reasonable. Definition 2 includes all valid kernels as well as some non-PSD non symmetric similarity functions [2, 3]. The authors have shown that this definition of good similarities allows also to solve problems that can not be handled by classical kernels, which makes the definition a strict generalization of kernels. According to the following theorem, it provides sufficient conditions to learn a good linear classifier in an explicit projection space defined by the reasonable points in the set R.

Theorem 1 (Balcan et al. [2]). *Let K be an $(\boldsymbol{\epsilon}, \boldsymbol{\gamma}, \boldsymbol{\tau})$-good similarity function for a learning problem P. Let $S = \{x'_1, \ldots, x'_d\}$ be a sample of $d = \frac{2}{\tau}\big(\log(\frac{2}{\delta}) + 8\frac{\log(2/\delta)}{\gamma^2}\big)$ landmarks (potentially unlabeled) drawn from P. Consider the mapping $\phi^R : X \to \mathbb{R}^d$ defined as follows: $\phi^R_i(x) = K(x, x'_i),\ i \in \{1, \ldots, d\}$. Then, with probability at least $1-\delta$ over the random sample R, the induced distribution $\phi^R(P)$ in $\mathbb{R}^d$ has a separator of error at most $\epsilon + \delta$ relative to L_1 margin at least $\gamma/2$.*

Thus, with an (ϵ,γ,τ)-good similarity function for a given learning problem P and enough (unlabeled) landmark examples, there exists with high probability a low-error linear separator in the explicit ϕ^R-space, corresponding to the space of the similarities to the d landmarks. The criterion given by Definition 2 requires to minimize the number of margin violations which is a NP-hard problem generally difficult to approximate. The authors have then proposed to consider an adaptation of Definition 2 with the hinge loss formalized as follows.

Definition 3 (Balcan et al. [2]). *A similarity function K is an* (ϵ, γ, τ)**-good similarity function in hinge loss** *for a learning problem P if there exists a (random) indicator function $R(x)$ defining a (probabilistic) set of "reasonable points" such that the following conditions hold:*

(i) we have

$$\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[[1 - yg(\mathbf{x})/\gamma]_+\Big] \leq \epsilon, \tag{2}$$

where $g(\mathbf{x}) = \mathbb{E}_{(\mathbf{x}',y')\sim P}[y'K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')]$
and $[1-c]_+ = \max(0, 1-c)$ *is the hinge loss,*

(ii) $Pr_{\mathbf{x}'}[R(\mathbf{x}')] \geq \tau$.

Using the same ϕ^R-space than Theorem 1, the authors have proved a similar theorem for this definition with the hinge loss. This leads to a natural two step algorithm for learning this classifier: select a set of potential landmark points and then learn a linear classifier in the projection space induced by these points. Then, using d_u unlabeled examples for the landmark points and d_l labeled examples, this linear separator $\boldsymbol{\alpha} \in \mathbb{R}^{d_u}$ can be found by solving a linear problem. We give here the formulation based on the hinge loss presented in [2].

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j y_i K(x_i, x'_j)\right]_+ \text{ such that} \sum_{j=1}^{d_u} |\alpha_j| \ \leq \ 1/\gamma. \tag{3}$$

In fact, we consider a similar formulation based on a 1-norm regularization, weighted by a parameter λ related to the desired margin.

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j y_i K(x_i, x'_j)\right]_+ + \lambda \|\boldsymbol{\alpha}\|_1. \tag{4}$$

In the following, a classifier learned in this framework is called a SF classifier.

4 Domain Adaptation

Domain adaptation (DA) [4, 20] arises when the learning data generation is somewhat different from the test data generation. The learning data, generally called the *source domain*, is represented by a distribution P_S over $X \times Y$ and the test data, referred to the *target domain*, is modeled by a distribution P_T. We denote by D_S and D_T the respective marginal distributions over X.
A learning algorithm is generally provided with a *Labeled Source sample* $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ drawn *i.i.d.* from P_S, and a *Target Sample* which contains a large set of unlabeled target instances $TS = \{\mathbf{x}_j\}_{j=1}^{m'}$ drawn *i.i.d.* from D_T and sometimes a few labeled target data drawn from P_T. The objective of a learning task is then to find a good hypothesis with a low error according to target distribution P_T. In this section, we provide a brief and non-exhaustive overview of some existing DA approaches, note that some surveys can be found in [18, 23]

The first theoretical analysis of the DA problem was proposed by Ben-David *et al.* [4, 5]. The authors have provided an upper bound on the target domain error err_{P_T} that takes into account the source domain error $\mathrm{err}_{P_S}(h)$ and the divergence $d_{\mathcal{H}}$ between the source and target marginal distributions:

$$\forall h \in \mathcal{H},\ \mathrm{err}_{P_T}(h) \ \leq\ \mathrm{err}_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu. \tag{5}$$

The last term corresponds to the optimal joint hypothesis over the two domains $\nu = \mathrm{argmin}_{h\in\mathcal{H}}\, \mathrm{err}_{P_S}(h) + \mathrm{err}_{P_T}(h)$. It can be seen as a quality measure of $\mathcal{H}$ for the DA problem considered. If this best hypothesis performs poorly, it appears then difficult to obtain a good hypothesis for the target domain. This term is then supposed to be small to ensure a successful adaptation.
The other crucial point is the divergence[2] $d_{\mathcal{H}}$ which is called the $\mathcal{H}$-distance. This result suggests that if the two distributions are close, then a low error classifier over the source domain can be a good classifier for the target one. The intuition behind this idea is given in Figure 1. The distance $d_{\mathcal{H}}$ is actually related to $\mathcal{H}$ by measuring a maximum variation divergence over the set of points on which an hypothesis in $\mathcal{H}$ can commit errors:

$$d_{\mathcal{H}}(D_S, D_T) \ =\ 2 \sup_{h\in\mathcal{H}} \left| Pr_{D_S}[I(h)] - Pr_{D_T}[I(h)] \right|$$

where $\mathbf{x} \in I(h) \Leftrightarrow h(\mathbf{x}) = 1$. An interesting point of this theory is that the $\mathcal{H}$-distance can be estimated from finite samples when the VC-dimension of $\mathcal{H}$ is finite. Using a VC-dimension analysis, the authors show that the empirical divergence converges to the true $d_{\mathcal{H}}$ with the size of the samples. Let U_S be a sample *i.i.d.* from D_S and U_T a sample *i.i.d.* from D_T. Consider a labeled sample $U_S \bigcup U_T$ where each instance of U_S is labeled as positive and each one of U_T as negative. The empirical divergence can then be directly estimated by looking for the best classifier able to separate the two samples[3] [4]:

$$\hat{d}_{\mathcal{H}}(U_S, U_T) \ =\ 2\left(1 - \min_{h\in\mathcal{H}} \hat{err}_{U_S,U_T}(h)\right), \tag{6}$$

with $\hat{err}_{U_S,U_T}(h) = \dfrac{1}{m}\left(\displaystyle\sum_{\substack{\mathbf{x}\in U_S\cup U_T,\\ h(\mathbf{x})=-1}} \mathbb{1}_{\mathbf{x}\in U_S} + \sum_{\substack{\mathbf{x}\in U_S\cup U_T,\\ h(\mathbf{x})=1}} \mathbb{1}_{\mathbf{x}\in U_T}\right)$, where $\mathbb{1}_{\mathbf{x}\in U_S} = \begin{cases} 1 \text{ if } \mathbf{x}\in U_S \\ 0 \text{ otherwise.} \end{cases}$

Note that finding the optimal hyperplane is NP-hard in general. However, a good estimation of $\hat{d}_{\mathcal{H}}$ allows us to have an insight of the distance between the two distributions and thus of the difficulty of the DA problem for the class $\mathcal{H}$. We will use this principle to estimate the difficulty of the task considered in our experimental part.

[2] The authors consider actually the divergence over $\mathcal{H}\Delta\mathcal{H}$, the space of symmetric difference hypothesis, see [4] for more details.

[3] By considering the 0-1 loss, L_{01}, defined as follows: $L_{01}(h, (\mathbf{x}, y)) = 1$ if $h(\mathbf{x}) \neq y$ and 0 otherwise.

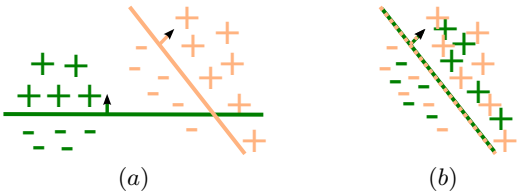


Fig. 1. Intuition behind a successful domain adaptation. Source points are in (dark) green (pos. +, neg. −), target points are in (light) orange. (*a*) The distance between domains is high: the two samples are easily separable and the classifier learned from source points performs badly on the target sample. (*b*) The distance between domains is low: The classifier learned from source points performs well on the two domains.

Later, Mansour *et al.* [20] have proposed another discrepancy measure allowing one to generalize the $d_{\mathcal{H}}$ distance to other real valued loss functions. Note that the bound presented in their work is a bit different from the one of Ben-David *et al.*. Moreover, they have also provided an average analysis with interesting Rademacher generalization bounds. These theoretical frameworks show that for a good domain adaptation, the distance between distributions and the source error must be low. According to [6], minimizing these two terms appears even necessary in general.

One key point for DA approaches is thus to be able to move closer the distributions while avoiding a dramatic increase of the error on the source domain. In the literature, some methods have proposed to reweight the source instances in order to get closer to the target distribution. They are often based on some assumptions on the two distributions [8, 17, 19, 20, 26]. For example some of these approaches rely on hypothesis like the covariate shift where the marginal distributions over X may be different for the two domains, but the conditional distribution of Y given X are the same, *i.e.* $P_S(y|\mathbf{x}) = P_T(y|\mathbf{x})$ for every $\mathbf{x} \in X$ and $y \in Y$ but $P_S(\mathbf{x}) \neq P_T(\mathbf{x})$ for some $\mathbf{x} \in X$ [26]. Other works are based on iterative self labeling approaches in order to move progressively from one domain to the other one [10]. Another standpoint for moving closer the two distributions is to find a relevant projection space where the two distributions are close. In [9], the authors propose a *structural correspondence learning* approach to identify relevant features by looking for their correspondence in the two domains. Another idea is to use an augmented feature space for both source and target data and use the new input space obtained with classical machine learning algorithms [11]. Some authors have also proposed to use spectral approaches to build a new feature space [21].

The main underlying ideas among the different approaches presented in this section is that a potential good adaptation needs to have the source and target distributions close. One way to achieve this goal is to build a relevant feature space by defining a new projection operator or by choosing relevant features. In the next section, we study the usefulness of the framework of Balcan *et al.*. to

deal with domain adaptation problems. More precisely, we propose to investigate how the definition of the similarity function and the construction of the feature space - *i.e.* the ϕ-space of similarities to a set of reasonable points - can help to improve the performance of the classifier in a domain adaptation setting.

5 Modifying the Projection Space for Domain Adaptation

In this section, we present our two approaches for modifying the similarity based projection space in order to facilitate the adaptation to the target distribution. First, we present a simple way for renormalizing a similarity function according to a sample of unlabeled instances. Second, we propose a regularization term that tends to define a projection space where the source and target marginal distributions tend to be closer.

5.1 A Normalization of a Similarity Function

For a particular DA task, we build a new similarity function K_N by normalizing a given similarity function K relatively to a sample N. Recall that, from Definition 2, a similarity must be good relatively to a set of reasonable points. We propose actually to renormalize the set of similarities to these points. Since the real set of reasonable points is unknown *a priori*, we consider a set of candidate landmark points R' and we apply a specific normalization for each instance of $\mathbf{x}'_j \in R'$. The idea is to apply a scaling to mean zero and standard deviation one for the similarities of the instances of N to $\mathbf{x}'$. Our procedure is defined as follows.

Definition 4. *Let K be a similarity function which verifies the Definition 2. Given a data set $N = \{\mathbf{x}_k\}_{k=1}^{p}$ and a set of (potential) reasonable points $R' = \{\mathbf{x}'_j\}_{j=1}^{d_u}$, a normalized similarity function, K_N, is defined by:*

$$\forall \mathbf{x}'_j \in R',\ K_N(.,\mathbf{x}'_j) = \begin{cases} \dfrac{K(.,\mathbf{x}'_j) - \mu_{\mathbf{x}'_j}}{\sigma_{\mathbf{x}'_j}} & \text{if } -1 \leq \frac{K(.,\mathbf{x}'_j)-\hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}} \leq 1, \\ -1 & \text{if } -1 \geq \frac{K(.,\mathbf{x}'_j)-\hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}} \\ 1 & \text{if } \frac{K(.,\mathbf{x}'_j)-\hat{\mu}_{\mathbf{x}'_j}}{\hat{\sigma}_{\mathbf{x}'_j}} \geq 1, \end{cases} \quad (7)$$

where $\hat{\mu}_{\mathbf{x}'_j}$ is the empirical mean of similarities to $\mathbf{x}'_j$ over N:

$$\forall \mathbf{x}'_j \in R',\ \hat{\mu}_{\mathbf{x}'_j} = \frac{1}{|N|} \sum_{\mathbf{x}_k \in N} K(\mathbf{x}_k, \mathbf{x}'_j),$$

and $\hat{\sigma}_{\mathbf{x}'_j}$ is the empirical unbiased estimate of the standard deviation:

$$\forall \mathbf{x}'_j \in R',\ \hat{\sigma}_{\mathbf{x}'_j} = \sqrt{\frac{1}{|N|-1} \sum_{\mathbf{x}_k \in N} \left(K(\mathbf{x}_k, \mathbf{x}'_j) - \hat{\mu}_{\mathbf{x}'_j}\right)^2}.$$

By construction, the similarity K_N is then non symmetric and non PSD. In the following, we will consider that a learning algorithm is provided with two data sets: $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ constituted of labeled source domain data, and $TS = \{\mathbf{x}_i\}_{i=1}^{m'}$ of unlabeled target domain data. According to the theoretical result of domain adaptation of Ben-David *et al.*. recalled in Equation (5), the learned classifier should also perform well on the source domain. We then propose to define our normalized function, denoted by K_{ST}, with $N = LS_{|X} \cup TS$ in order to link the two domains by considering the information of both of them at the same time, for avoiding an increasing of the source error. Our choice is clearly heuristic and our aim is just to evaluate the interest of renormalizing a similarity for domain adaptation problems. In order to study the potential of adaptation, we will only consider candidate landmark points R' from the source domain.

5.2 An Additional Regularization Term For Moving Closer the Two Distributions

As a second contribution, we propose to add a regularization term to the optimization Problem 4 proposed by Balcan *et al.*. The objective is to control the selection of reasonable points leading to a projection space where the two distributions are close. According to the empirical divergence $d_{\mathcal{H}}$ given in Equation 6, the source and the target domains are close if it is difficult to separate source from target examples. Let two subsets $U_S \subseteq LS$ and $U_T \subseteq TS$ of equal size, our idea is then to build a set $\mathcal{C}_{ST}$ of pairs belonging to $U_S \times U_T$. And then, for each pair $(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}$, we propose to regularize the learned classifier such that the outputs of the classifier are close for the two instances $\mathbf{x}_s$ and $\mathbf{x}_t$. For any classifier $h(\cdot) = \sum_{i=1}^{|R|} \alpha_i K(\cdot, x'_i)$, this can be expressed as follows:

$$
\begin{aligned}
|h(\mathbf{x}_s) - h(\mathbf{x}_t)| &= \left| \sum_{j=1}^{|R|} \alpha_j K(\mathbf{x}_s, \mathbf{x}'_j) - \sum_{j=1}^{|R|} \alpha_j K(\mathbf{x}_t, \mathbf{x}'_j) \right| \\
&\leq \sum_{j=1}^{|R|} \left| \alpha_j \big(K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j) \big) \right| \text{ by using triangle inequality} \\
&= \left\| \left({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t) \right) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1 . \qquad (8)
\end{aligned}
$$

This leads us to propose a new regularization term which tends to select landmarks with similarities close to both some source and target points, which allows us to define a projection space where source and target examples are closer. Let R be a set of d_u candidate landmark points, our global optimization problem is then defined as follows:

$$
\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j y_i K(x_i, x'_j) \right]_+ + \lambda \|\boldsymbol{\alpha}\|_1 + C \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} \left\| \left({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t) \right) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1 \qquad (9)
$$

The construction of C_{ST} is difficult since we have no information on the target labels. In practice, we build the matching C_{ST} from U_S and U_T by looking for a bipartite matching minimizing the Euclidean distance in the ϕ-space defined by the set of candidate landmarks. This can be done by solving the following problem. Note that in the particular case of bipartite matching, this can be done in polynomial time by linear programming for example.

$$\begin{cases} \min\limits_{\substack{\beta_{st} \\ 1\leq s\leq |U_S| \\ 1\leq t\leq |U_T|}} \sum\limits_{(\mathbf{x}_s,\mathbf{x}_t)\in U_S\times U_T} \beta_{st}\|\phi^R(\mathbf{x}_s)-\phi^R(\mathbf{x}_t)\|_2^2 \\ \text{s.t.: } \forall (\mathbf{x}_s,\mathbf{x}_t)\in U_S\times U_T,\ \beta_{st}\in\{0,1\}, \\ \qquad\qquad \forall \mathbf{x}_s\in U_S,\ \sum_{\mathbf{x}_t\in U_T}\beta_{(st)}=1, \\ \qquad\qquad \forall \mathbf{x}_t\in U_T,\ \sum_{\mathbf{x}_s\in U_S}\beta_{(st)}\leq 1. \end{cases}$$

Then $\mathcal{C}_{ST}$ corresponds to the pairs of $U_S \times U_T$ such that $\beta_{st} = 1$. The choice of the points of U_S and U_T is hard and in an ideal case, we would like to select pairs of points of the same label. But since we suppose that no target label is available, we select the sets U_S and U_T randomly from the source and target samples, from different draws, and we choose the best sets thanks to a reverse validation procedure described in Appendix A.

6 Experiments

We now propose to evaluate the approaches presented in the previous section on a synthetic toy problem and on a real image annotation task. For every problem, we consider to have: a labeled source sample LS drawn from the source domain, a set of potential landmark points R' drawn from the marginal source distribution over X and an unlabeled target sample TS drawn from the marginal target distribution over X.
As a baseline, we choose a similarity based on a classical Gaussian kernel, which is a *good similarity function* according to the framework of Balcan *et al.*:

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{D^2}\right).$$

We then consider the normalized similarity K_{ST} which corresponds to the normalization of K according to the instances of the source and target samples $LS_{|X} \cup TS$. For each of the two similarities K and K_{ST}, we compare the models learned by solving Problem (4), corresponding to learning a classical SF-based classifier, to those learned using our regularized formulation in Problem (9). We tune the hyperparameters with a "reverse" validation procedure described in Appendix A. Moreover, in order to evaluate if K_{ST} is a better similarity for the target domain, we propose to study the (ϵ,γ,τ)-guarantees on the target sample according to Definition 3. For this purpose, we estimate empirically ϵ as a function of γ from the target sample (we use here the real labels but only for this

evaluation), *i.e.* for a given γ, $\hat{\epsilon}$ is the proportion of examples $\mathbf{x} \in TS$ verifying:

$$\sum_{\mathbf{x}'_j \in R'} y_i y'_j K(\mathbf{x}_i, \mathbf{x}'_j) < \gamma.$$

We also assess the distance $\hat{d}_{\mathcal{H}}$ between the two domains by learning a SF-based classifier with K for separating source from target samples in the original space. From Equation (6), a small value, near 0, indicates close domains while a larger value, near 2, indicates a hard DA task.

6.1 Synthetic Toy Problem

As the source domain, we consider a classical binary problem with two intertwinning moons, each class corresponding to one moon (see Figure 3). We then define 8 different target domains by rotating anticlockwise the source domain according to 8 angles. The higher the angle is, the harder the task becomes. For each domain, we generate 300 instances (150 of each class). Moreover, for studying the influence of the pair set $\mathcal{C}_{ST}$, we evaluate the obtained results when $\mathcal{C}_{ST}$ corresponds to a set of "perfect pairs $(\mathbf{x}_s, \mathbf{x}_t)$" where $\mathbf{x}_t$ is the obtained instance after rotating $\mathbf{x}_s$. These results correspond to an upper bound for our methods. Finally, in order to assess the generalization ability of our approach, we evaluate each method on an independent test set of 1500 examples drawn from the target domain (not provided to the algorithm). Each adaptation problem is repeated 10 times and the average accuracy obtained for each method is reported in Table 1. We can make the following remarks.

- Our new regularization term for minimizing distance between marginal distributions improves significantly the performances on the target domain.
- As long as the problem can be considered as an easy DA task, the normalized similarity does not produce better models. However, when the difficulty increases, using a normalized similarity improves the results.
- Regarding the bipartite matching influence, having perfect pairs leads to the best results and is thus important for the adaption process, which is expected. However, our reverse validation procedure helps us to keep correct results when a set of perfect pairs can not be built.

Figure 2 shows the goodness guarantees of the similarities over each adaptation task. A better similarity has a lower area under the curve, meaning a lower error in average. The $\hat{\epsilon}$ rate is relatively high because we consider only landmarks from the source sample in order to study our adaptation capability. We observe for hardest problems ($\geq 50°$) an improvement of the goodness with the normalized similarity K_{ST}. For easier tasks, this improvement is not significant, justifying the fact that the similarity K can lead to better classifiers. Our normalized similarity seems thus relevant only for hard domain adaptation problems.

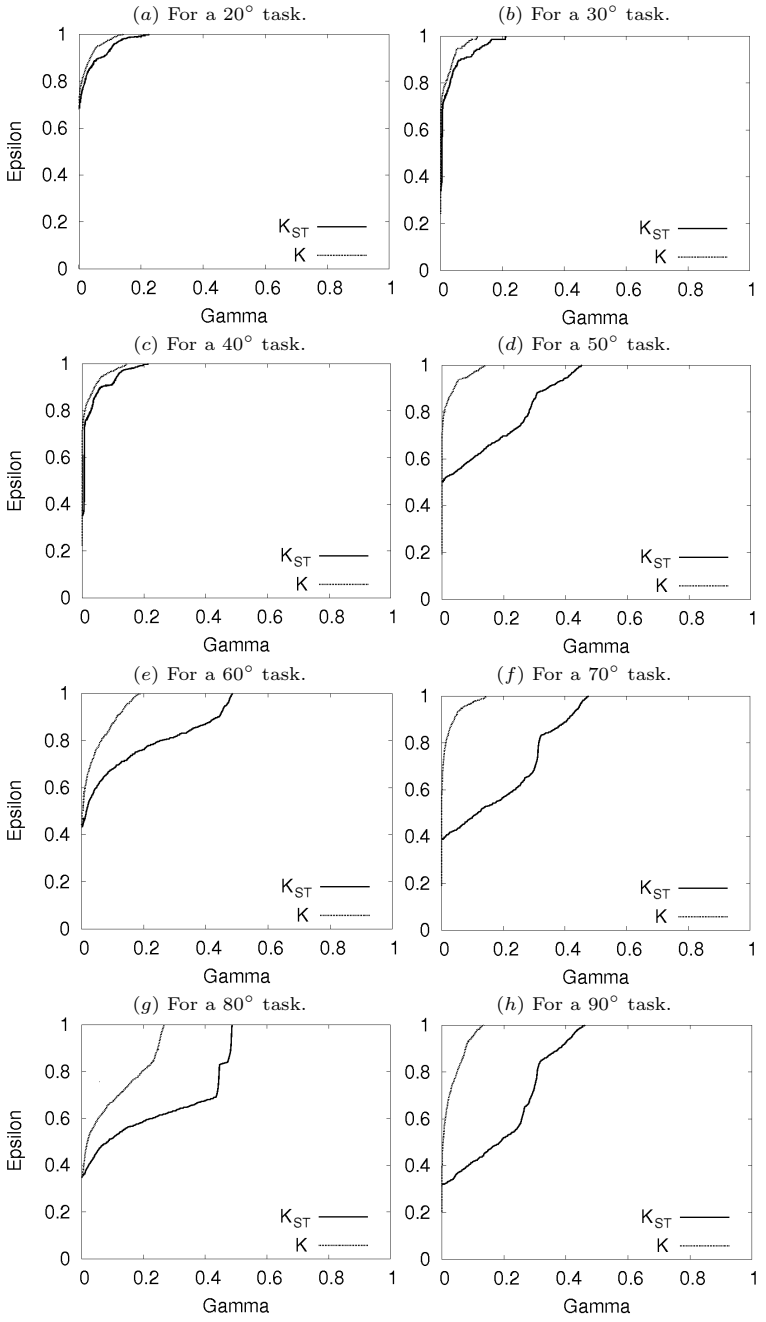


Fig. 2. Goodness of the similarities over the target sample: $\hat{\epsilon}$ as a function of γ.

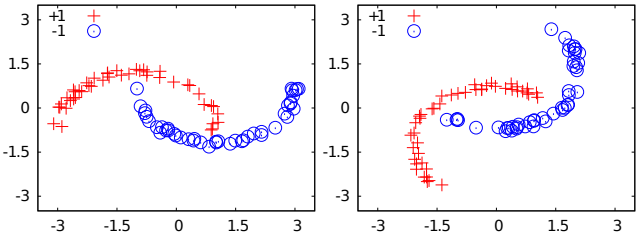


Fig. 3. Left: A source sample. Right: A target sample with a 50° rotation.

ROTATION	20°	30°	40°	50°	60°	70°	80°	90°
$\hat{\mathbf{d}}_{\mathcal{H}}$	0.58	1.16	1.31	1.34	1.34	1.32	1.33	1.31
SF without distance regularization								
WITH K	88 ± 13	70 ± 20	59 ± 23	47 ± 17	34 ± 08	23 ± 01	21 ± 01	19 ± 01
WITH K_{ST}	79 ± 10	56 ± 15	56 ± 10	43 ± 09	41 ± 08	37 ± 10	36 ± 10	40 ± 09
SF with distance regularization								
WITH K	**98** ± 03	**92** ± 07	**83** ± 05	70 ± 09	54 ± 18	43 ± 24	38 ± 23	35 ± 19
WITH K_{ST}	93 ± 05	86 ± 08	72 ± 12	**72** ± 013	**69** ± 10	**67** ± 12	**63** ± 13	**58** ± 09
SF with distance regularization and perfect matching								
WITH K	***99*** ± 01	***96*** ± 01	***86*** ± 02	73 ± 11	65 ± 23	56 ± 29	47 ± 23	39 ± 19
WITH K_{ST}	97 ± 04	92 ± 06	83 ± 10	***75*** ± 12	***73*** ± 16	***73*** ± 02	***69*** ± 7	***60*** ± 11

Table 1. Average results in percentage of accuracy with standard deviation on the toy problem target test sample for each method.

6.2 Image Classification

In this section, we experiment our approach on PascalVOC 2007 [13] and TrecVid 2007 [25] corpora. The PascalVOC benchmark is constituted of a set of 5000 training images and a set of 5000 test images. The TrecVid corpus is constituted of images extracted from videos and can be seen also as an image corpus. The goal is to identify visual objects and scenes in images and videos. We choose the concepts that are shared between the two corpora: `Boat`, `Bus`, `Car`, `TV/Monitor`, `Person` and `Plane`. We used visual features extracted as described in [1]. We consider as the source domain, labeled images from the PascalVOC 2007 training set. For each concept, we generated a source sample constituted of all the training positive images and negatives images independently drawn such that the ratio $+/-$ is $\frac{1}{3}/\frac{2}{3}$. As the target domain, we use some images of the TrecVid corpus, we built also a sample containing all the positive examples and drew some negative samples in order to keep the same ration $+/-$ of $\frac{1}{3}/\frac{2}{3}$. In these samples, the number of positive examples may be low and we propose to use the *F-measure*[4] to evaluate the learned models. The results are reported in Table 2. The different nature and ways of acquisition of the images make the problem of adaptation difficult. As an illustration, the empirical $\hat{d}_{\mathcal{H}}$ between the two domains is high for every concept. In this context, for all the tasks, the normalized similarity with

[4] The F-measure or the balanced F-score is the harmonic mean of precision and recall.

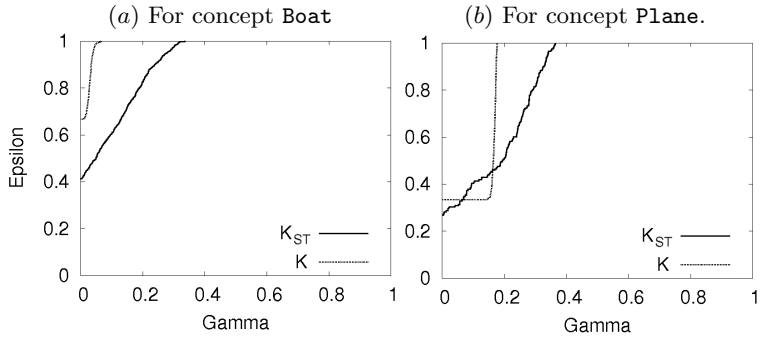


Fig. 4. $\hat{\epsilon}$ on the target domain as a function of γ for 2 concepts.

CONCEPT	BOAT	BUS	CAR	MONITOR	PERSON	PLANE	AVERAGE
$\hat{\mathbf{d}}_{\mathcal{H}}$	1.93	1.95	1.85	1.86	1.78	1.86	1.86
SF without distance regularization							
WITH K	0.0279	0.1806	0.5214	0.2477	0.4971	0.5522	0.3378
WITH K_{ST}	0.4731	0.4632	0.5316	0.3664	0.3776	0.5635	0.4626
SF with distance regularization							
WITH K	0.2006	0.1739	0.5125	0.2744	0.5037	0.5192	0.3640
WITH K_{ST}	**0.4857**	**0.4891**	**0.5452**	**0.3989**	**0.5353**	**0.6375**	**0.5153**

Table 2. Results obtained on TrecVid target domain according to the F-measure.

distance regularization provides the best results. This is confirmed on Figure 4 where the evaluation of the goodness of the two similarities for two concepts is provided: the normalized similarity is better for difficult tasks.

7 Conclusion

In this paper, we have proposed a preliminary study on the usefulness of the framework of Balcan *et al.* [2, 3] for domain adaptation. We have proposed a normalization of a similarity function according to a test sample based on the fact that a similarity does not need to be PSD or symmetric. We have also proposed a new regularization term that tends to define a projection space of reasonable points where the source and target distributions of the examples are closer. We have provided experiments on a toy problem and on a real image annotation task. Our regularization term generally helps to improve the learned classifier and the normalization proposed seems only relevant for difficult adaptation tasks.

As a future work, we will continue on the idea of normalizing a similarity in order to adapt it to the target domain. Around this idea, many questions remain open like the choice the landmark points, the influence of the test set or avoiding overfitting. The use of some labeled target data may also help to produce a better projection space. From a theoretical standpoint, a perspective

would be to consider an extension of the framework of robustness of Xu and Mannor [28] to domain adaptation.

References

1. Ayache, S., Quénot, G., Gensel, J.: Image and video indexing using networks of operators. Journal on Image and Video Processing 2007, 1:1–1:13 (2007)
2. Balcan, M.F., Blum, A., Srebro, N.: Improved guarantees for learning via similarity functions. In: Proceedings of COLT. pp. 287–298 (2008)
3. Balcan, M.F., Blum, A., Srebro, N.: A theory of learning with similarity functions. Machine Learning Journal 72(1-2), 89–112 (2008)
4. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning Journal 79(1-2), 151–175 (2010)
5. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: Proceedings of NIPS'06. pp. 137–144 (2006)
6. Ben-David, S., Lu, T., Luu, T., Pal, D.: Impossibility theorems for domain adaptation. JMLR W&CP 9, 129–136 (2010)
7. Bernard, M., Boyer, L., Habrard, A., Sebban, M.: Learning probabilistic models of tree edit distance. Pattern Recognition 41(8), 2611–2629 (2008)
8. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning for differing training and test distributions. In: Proceeding of ICML. pp. 81–88 (2007)
9. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of EMNLP. pp. 120–128 (2006)
10. Bruzzone, L., Marconcini, M.: Domain adaptation problems: A DASVM classification technique and a circular validation strategy. IEEE Trans. Pattern Anal. Mach. Intell. 32(5), 770–787 (2010)
11. Daumé III, H.: Frustratingly easy domain adaptation. In: Proceedings of the Association for Computational Linguistics (ACL) (2007)
12. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: Proceedings of ICML. pp. 209–216 (2007)
13. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. www.pascal-network.org/challenges/VOC/voc2007/workshop/ (2007)
14. Gao, X., Xiao, B., Tao, D., Li, X.: A survey of graph edit distance. Pattern Analysis & Applications 13(1), 113–129 (2010)
15. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Proceedings of NIPS. vol. 17, pp. 513–520 (2004)
16. Haasdonk, B.: Feature space interpretation of svms with indefinite kernels. IEEE Trans. Pattern Anal. Mach. Intell. 27(4), 482–492 (2005)
17. Huang, J., Smola, A., Gretton, A., Borgwardt, K., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: Proceedings of NIPS. pp. 601–608 (2006)
18. Jiang, J.: A literature survey on domain adaptation of statistical classifiers. Tech. rep., Computer Science Department at University of Illinois at Urbana-Champaign (2008), http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/da_survey.pdf
19. Jiang, J., Zhai, C.: Instance weighting for domain adaptation in NLP. In: Proceedings of ACL (2007)
20. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation: Learning bounds and algorithms. In: Proceedings of COLT. pp. 19–30 (2009)

21. Pan, S., Tsang, I., Kwok, J., Yang, Q.: Domain adaptation via transfer component analysis. In: Proceedings of IJCAI. pp. 1187–1192 (2009)
22. Pan, S., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
23. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Dataset Shift in Machine Learning. The MIT Press (2009)
24. Ristad, E., Yianilos, P.: Learning string-edit distance. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(5), 522–532 (1998)
25. Smeaton, A., Over, P., Kraaij, W.: High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In: Multimedia Content Analysis, Theory and Applications, pp. 151–174. Springer Verlag, Berlin (2009)
26. Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M.: Direct importance estimation with model selection and its application to covariate shift adaptation. In: Proceedings of NIPS (2007)
27. Weinberger, K., Saul, L.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research (JMLR) 10, 207–244 (2009)
28. Xu, H., Mannor, S.: Robustness and generalization. In: Proceedings of COLT. pp. 503–515 (2010)
29. Zhong, E., Fan, W., Yang, Q., Verscheure, O., Ren, J.: Cross validation framework to choose amongst models and datasets for transfer learning. In: Proceedings of ECML-PKDD (Part III). LNCS, vol. 6323, pp. 547–562. Springer (2010)

A Appendix

Given a classifier h, we define the *reverse classifier* h^r as the classifier learned from the target sample self labeled by $h : \{(\mathbf{x}, \text{sign}(h(\mathbf{x})))\}_{\mathbf{x}\in TS}$. According to the idea of Zhong *et al.* [10, 29], we evaluate h^r on the source domain (see Fig. 5). Given k-folds on the source labeled sample, we use $k-1$ folds as labeled examples for solving Pb. (9) and we evaluate h^r on the last k^{th} fold. The final error corresponds to the mean of the error over the k-folds: $\hat{\text{err}}_S(h^r) = \frac{1}{k}\sum_{i=1}^{k} \hat{err}_{LS_i}(h^r)$. Among many classifiers h, the one with the lowest $\hat{\text{err}}_S(h^r)$ is chosen.

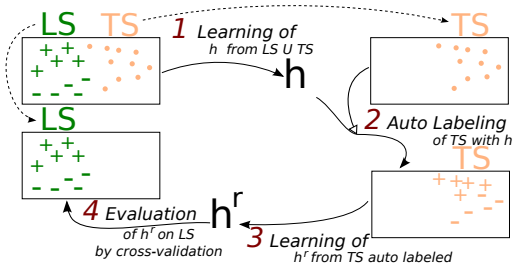


Fig. 5. Reverse validation. Step **1**: *Learning* h with Problem (9). **2**: *Auto-labeling* the target sample with h. **3**: *Learning* h^r on auto-labeled target sample by Problem (4). **4**: *Evaluation* of h^r on LS (with a k-folds process) for validating h.