LARGE DIMENSIONAL ANALYSIS OF LS-SVM TRANSFER LEARNING: APPLICATION TO POLSAR CLASSIFICATION

Cyprien Doz[1], *Chengfang Ren*[1], *Jean-Philippe Ovarlez*[1,3], *Romain Couillet*[2],

[1]SONDRA, CentraleSupélec, University of Paris-Saclay
[2]LIG-Lab, University of Grenoble-Alpes,
[3]DEMR, ONERA, University of Paris-Saclay

ABSTRACT

This article analyzes a kernel-based transfer learning method, under a k-class Gaussian mixture model for the input data. Following recent advances in random matrix theory, we propose new insights in transfer learning schemes for challenging cases, when the first-order statistics of all data classes coincide. The article proves the asymptotic normality of the LS-SVM decision function for any smooth kernel function. As a result, an optimization scheme is proposed to minimize the classification error rate. Our theoretical results are corroborated through simulations and then successfully applied to the context of transfer learning for PolSAR image classification.

Index Terms— Transfer learning, high dimensional statistics, kernel methods, random matrix theory, support vector machines.

1. INTRODUCTION

In classical machine learning, tasks are generally processed separately. This approach however does not take into account the potentially high similarity between tasks. Transfer learning aims to leverage information contained in one task (the source task) to help improve the generalization performance of another task (the target task); see [1] and [2] for detailed tutorials and various interpretations of the actual methods.

This article specifically focuses on a *least-square support vector machine* (LS-SVM) version of transfer learning, as introduced in [3]. The approach followed by [3] is to share part of the separating hyperplane in both source and target tasks and then solve a parallel SVM optimization for both tasks, under this shared hyperplane constraint. Despite the simplicity of the method, the comprehension of the transfer learning behavior and performances remains quite empirical. Particularly fundamental is the question of the appropriate choice of the SVM hyperparameters in order to avoid the dreaded problem of *negative transfer*, by which the source task optimization actually impedes rather than helps the target task.

This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003)

First breakthroughs were recently made in [4,5] by assuming a large dimensional statistical model, using modern random matrix advances. The present article leverages these recent findings to generalize the LS-SVM model to a *kernel LS-SVM* transfer learning mechanism, naturally appropriate to handle complex data structures. Specifically, the article provides a theoretical analysis generalizing [4,5] to a data model with different covariance structures for each data class. As in [4,6], our analysis reveals the importance of optimizing the training data labels in order to combat negative transfer.

Our main contributions may be summarized as follows: (i) we provide a theoretical estimate of the limiting performance of LS-SVM transfer learning under a large and numerous data assumption; (ii) we analyze the relation between covariance matrix structures in all data classes (source and target) and the transfer learning inner workings; (iii) these results are applied to the practical setting of PolSAR classification.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices), and scalars non-boldface respectively. $\mathbf{1}_n$ is the column vector containing n ones, and $\mathbf{I}_n$ the $n \times n$ identity matrix. The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices. $\mathbf{P}$ and $\mathbf{P}_c$ are centering matrices used to normalize the data and suppress biases inherent to theoretical analysis.

2. MODEL AND ASSUMPTIONS

2.1. Model

Consider $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ a set of n samples arising from k classes, with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{\ell_1, \ldots, \ell_k\}$ for $\ell_a \in \mathbb{R}$ the label associated to class $\mathcal{C}_a$. We define $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}$ and $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_k]^\mathsf{T}$. Note that $\mathbf{y} = \mathbf{J}\,\boldsymbol{\ell}$ with, $\mathbf{J} \triangleq [\mathbf{j}_1, \ldots, \mathbf{j}_k]$, where $\mathbf{j}_a = \{\delta(y_i = \ell_a)\}_{i=1}^n$ is the indicator vector[1] of class $\mathcal{C}_a$ with cardinality n_a.

We further denote $c_a = n_a/n$ for $a \in \{1, \ldots, k\}$, $\mathbf{c} = [c_1, \ldots, c_k]^\mathsf{T}$ and $\mathbf{P} \triangleq \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$, $\mathbf{P}_c \triangleq \frac{1}{n}\,\mathbf{J}^\mathsf{T}\,\mathbf{P}\,\mathbf{J} =$

[1]where $\delta(.)$ is the *Kronecker symbol*

$\mathrm{diag}(\mathbf{c}) - \mathbf{c}\,\mathbf{c}^\mathsf{T}$.

The set of data is divided between a *target* subset, which contains samples from exclusively two classes (referred to as target classes), and a *source* subset, which contains all other samples, often numerous, compared to the target subset.

Our practical problem is to classify unknown new data into one of the two target classes while benefiting from the existence of the labeled source data.

The task of interest is thus the classification into classes $\mathcal{C}_{T_1}$ and $\mathcal{C}_{T_2}$, with $T_1, T_2 \in \{1, \ldots, k\}$, while training on all samples available, i.e., including the source data to the training process. In order to simplify the analysis, we focus, without loss of generality, on the setting of $k = 4$ classes (in the following, we will denote these class indexes by S_1, T_1, S_2, T_2): $\mathcal{C}_{T_1}$ and $\mathcal{C}_{T_2}$ on the target side and $\mathcal{C}_{S_1}$ and $\mathcal{C}_{S_2}$ on the source side. The extension to a multi-class scenario is however immediate. One could imagine a scenario with k_T target classes and k_S source classes, by imagining an extension *one versus all* of the binary case.

The classification performance on $\mathcal{C}_{T_1}$ and $\mathcal{C}_{T_2}$ will naturally strongly depend on the distribution similarity of data in source classes. This is here encapsulated in the kernel matrix $\mathbf{K}$, with entries $\mathbf{K}_{ij} = f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)$, for $f : \mathbb{R} \to \mathbb{R}$, where f is the kernel function.

2.2. Problem statement

LS-SVM [7] aims to predict the class label $\ell_{\mathbf{x}}$ of incoming data $\mathbf{x}$, thanks to a training performed on the training dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, by devising a separating hyperplane $\mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}) + b$ between $\mathcal{C}_{T_1}$ and $\mathcal{C}_{T_2}$, which is defined by the optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\arg\min} \quad & L(\mathbf{w}, e) = \|\mathbf{w}\|^2 + \frac{\gamma}{n}\sum_{i=1}^n e_i^2 \\ \text{such that} \quad & e_i = y_i - \mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_i) - b,\ i = 1, \ldots, n \end{aligned} \tag{1}$$

where $\gamma > 0$ is the penalty factor that balances the risk due to the potential complexity of the model, $\|\mathbf{w}\|^2$, and the risk due to the distance between real labels and estimated ones. The resolution (as in [8], p.4) of (1) by the method of Lagrange multiplier $\boldsymbol{\alpha}$ leads to the solution $\hat{\mathbf{w}} = [\boldsymbol{\varphi}(\mathbf{x}_1)\ \cdots\ \boldsymbol{\varphi}(\mathbf{x}_n)]^\mathsf{T}\boldsymbol{\alpha} = \sum_{i=1}^n \alpha_i \boldsymbol{\varphi}(\mathbf{x}_i)$, where

$$\begin{cases} \boldsymbol{\alpha} &= \mathbf{Q}\left(\mathbf{I}_n - \dfrac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\mathbf{Q}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\,\mathbf{1}_n}\right)\mathbf{y} = \mathbf{Q}\,(\mathbf{y} - b\,\mathbf{1}_n) \\ b &= \dfrac{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\,\mathbf{y}}{\mathbf{1}_n^\mathsf{T}\mathbf{Q}\,\mathbf{1}_n} \end{cases} \tag{2}$$

with $\mathbf{Q} = \left(\mathbf{K} + \dfrac{n}{\gamma}\mathbf{I}_n\right)^{-1}$ and $\mathbf{K} = \left\{\boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j)\right\}_{i,j}$.
Using the *kernel trick*, we now denote $\mathbf{K} = \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$.

Given $\boldsymbol{\alpha}$ and b, a new datum $\mathbf{x}$ is then classified into classes $\mathcal{C}_a$ depending on the value of the decision function:

$$g(\mathbf{x}) = \sum_{i=1}^n \alpha_i \boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}) + b = \boldsymbol{\alpha}^\mathsf{T}\mathbf{k}(\mathbf{x}) + b\,. \tag{3}$$

where $\mathbf{k}(\mathbf{x}) = \{f(\|\mathbf{x} - \mathbf{x}_j\|^2/p)\}_{j=1}^n$. A new datum $\mathbf{x}$ is associated with class, say, $\mathcal{C}_{T_1}$ if $g(\mathbf{x})$ is below a certain threshold, and with $\mathcal{C}_{T_2}$ otherwise. To classify *target* data, the decision threshold of $g(\mathbf{x})$ is to be defined and we will define it in the section 3 based on its asymptotic distribution.

In high dimension, most machine learning methods, including LS-SVM, consider that the algorithms work in a regime $p << n$. Before proceeding to the LS-SVM transfer learning main results, a few extra technical assumptions are needed.

2.3. Assumptions and non-trivial regime

We suppose that, for $a \in \{1, \ldots, k\}$:

$$\mathbf{x}_i \in \mathcal{C}_a \quad \text{if} \quad \mathbf{x}_i = \boldsymbol{\mu}_a + \sqrt{p}\,\boldsymbol{\omega}_i\,,$$

where $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, p^{-1}\,\mathbf{C}_a)$ with $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ a symmetric and non-negative definite matrix. In this paper, we will focus on the non-trivial problem of separating centered data, i.e. $\boldsymbol{\mu} = \mathbf{0}$.
To avoid trivial results, we assume, as in [5,9], that the class statistics satisfy certain growth rate conditions between the data covariance matrices. Setting $\mathbf{C}^\circ \triangleq \dfrac{1}{n}\sum_{a=1}^k n_a\,\mathbf{C}_a$ and defining the parameter $\tau \triangleq \dfrac{2}{p}\,\mathrm{tr}\,\mathbf{C}^\circ > 0$, it was shown in [9] that:

$$p^{-1}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \overset{\text{a.s.}}{\to} 0 \quad \text{for any } i \neq j, \tag{4}$$

This (problematic) phenomenon of concentration of distance allows us to approximate asymptotically $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ by a Taylor expansion (of order two) of f around τ. We will assume here that the kernel function f is at least three times derivable in a neighborhood of τ. Thus, the asymptotic behavior of LS-SVM transfer learning will strongly depend on the first two derivatives of f around τ.

As such, the asymptotic behavior of LS-SVM transfer learning will heavily depend on the first two derivatives of f in τ.

LS-SVM classification performances in high dimensions will heavily depend on the nature of the kernel function. This feature becomes adaptable to different given problems as explored in [9].

In the classical statistic setting, the features of SVM framework have made the analysis of its performances a challenging task. SVMs performances have been studied with various approaches (introducing VC dimension [8] or Bayesian interpretation [10]) keeping p and n fixed. The

recent asymptotic analysis conducted in [5] opened the door to tracks of improvement for LS-SVM framework. Now, we intend to conduct the same kind of asymptotic analysis in order to explore the inner-working between source and target data in a transfer learning context. LS-SVM classification performances in high dimensions will heavily depend on the nature of the kernel function. This feature becomes adaptable to different given problem as explored in [9] and [6] for vanishing difference in means across classes.

With the previous assumptions at hand, we are now capable of running through a technical analysis of our LS-SVM transfer learning problem in high dimensions.

3. ASYMPTOTIC RESULTS

In order to assess the performance of the classifier and the impact of source data in the training set, as $n, p \to \infty$, our objective is to provide an approximation of $g(\mathbf{x})$ in this regime. The solution $(\boldsymbol{\alpha}, b)$ of (2) is a function of $\mathbf{y}$ and of $\mathbf{Q}$. As $\mathbf{Q}$ is not directly accessible, we proceed in two steps: (i) exploiting the results of [9], we exploit a technically convenient asymptotic "linearization" of $\mathbf{K}$, (ii) performing a Taylor-expansion on $\mathbf{Q}$ around its dominant term $\left(f(\tau)\,\mathbf{1}_n\,\mathbf{1}_n^\mathsf{T} + \frac{n}{\gamma}\,\mathbf{I}_n\right)^{-1}$. Proceeding as in [5], we then obtain asymptotically accurate approximations for $\boldsymbol{\alpha}$ and b. In the case of applications of the paper, we focus on the classification of data with zero mean vector (and thus only discriminated by class co-variances). In line with remarks made in [5], it is more interesting to use kernels verifying $f'(\tau) = 0$. By setting:

$$
\begin{aligned}
t_a &= p^{-1/2}\,\mathrm{tr}(\mathbf{C}_a - \mathbf{C}^\circ),\ \ \mathbf{t} = [t_1, \ldots, t_k]^T\,, && (5)\\
\mathbf{t}_{\mathbf{C},a} &= [\mathrm{tr}(\mathbf{C}_a\,\mathbf{C}_1), \ldots, \mathrm{tr}(\mathbf{C}_a\,\mathbf{C}_k)]^\mathsf{T}\,, && (6)\\
\mathbf{T} &= [\mathbf{t}_{\mathbf{C},1}, \ldots, \mathbf{t}_{\mathbf{C},k}]\,, && (7)\\
\mathfrak{m}_a &= p^{-1}\,f''(\tau)\,\mathbf{t}\,t_a + 2\,p^{-2}\,f''(\tau)\,\mathbf{t}_{\mathbf{C},a}\,, && (8)\\
\mathbf{W}_a &= p^{-3}\,\left(f''(\tau)\right)^2\,\mathbf{t}\,\mathbf{t}^\mathsf{T}\,\mathrm{tr}\,\mathbf{C}_a^2\,, && (9)\\
C_\mathbf{T} &= -2\,p^{-2}\,f''(\tau)\,\mathbf{c}^\mathsf{T}\,\mathbf{T}\,\mathbf{P}_c\,\boldsymbol{\ell}\,, && (10)
\end{aligned}
$$

the main result of [5] extends as follows.

Theorem 1 (Gaussian Approximation)**.** *Let* $\mathbf{x} \in \mathcal{C}_a, a \in \{T_1, T_2\}$. *Considering that the stated hypotheses are verified, the asymptotic distribution of* $g(\mathbf{x})$ *defined by* (3) *is given by:*

$$
n\,\mathrm{V}_a^{-\frac{1}{2}}\left(g\left(\mathbf{x} \mid \mathbf{x} \in \mathcal{C}_a\right) - \mathrm{E}_a\right) \xrightarrow{d} \mathcal{N}(0,1) \qquad (11)
$$

where mean E_a *and variance* V_a *are defined as*

$$
\begin{aligned}
\mathrm{E}_a &= \mathbf{c}^\mathsf{T}\boldsymbol{\ell} + \gamma\,\boldsymbol{\ell}^\mathsf{T}\,\mathbf{P}_c\,\mathfrak{m}_a + \gamma\,C_\mathbf{T}\,, && (12)\\
\mathrm{V}_a &= 2\gamma^2\boldsymbol{\ell}^\mathsf{T}\mathbf{P}_c\,\mathbf{W}_a\mathbf{P}_c\boldsymbol{\ell}\,, && (13)
\end{aligned}
$$

This theorem allows us to characterize the classification performance for the two classes. An example is illustrated through the histogram at the left-hand side of Fig. 2 for the Toy-example of Section 4.1. Thus, we can also characterize the optimal decision threshold that separates the classes $\mathcal{C}_{T_1}$ and $\mathcal{C}_{T_2}$. If we denote by s the decision threshold for class membership, the classification error P_e is given by: $g(\mathbf{x}) \underset{\mathcal{C}_{T_1}}{\overset{\mathcal{C}_{T_2}}{\gtrless}} s$. With this value of s fixed, the classification error rate P_e is given by:

$$
P_e = \frac{1}{2}\left(P\left(g(\mathbf{x}) > s \mid \mathbf{x} \in \mathcal{C}_{T_1}\right) + P\left(g(\mathbf{x}) < s \mid \mathbf{x} \in \mathcal{C}_{T_2}\right)\right).
$$

With the result of Theorem 1 and after some manipulations[2], we can define the optimal classification threshold[3] s_{opt} between two Gaussian variables $G_1 \sim \mathcal{N}(\mathrm{E}_{\mathrm{T}_1}, \mathrm{V}_{\mathrm{T}_1})$ and $G_2 \sim \mathcal{N}(\mathrm{E}_{\mathrm{T}_2}, \mathrm{V}_{\mathrm{T}_2})$ minimizing probability P_e :

$$
\begin{aligned}
s_{opt} = &\frac{\mathrm{E}_{\mathrm{T}_2}\mathrm{V}_{\mathrm{T}_1} - \mathrm{E}_{\mathrm{T}_1}\mathrm{V}_{\mathrm{T}_2}}{\mathrm{V}_{\mathrm{T}_1} - \mathrm{V}_{\mathrm{T}_2}} - \frac{(\mathrm{V}_{\mathrm{T}_1}\mathrm{V}_{\mathrm{T}_2})^{1/2}}{\mathrm{V}_{\mathrm{T}_1} - \mathrm{V}_{\mathrm{T}_2}}\times\\
&\left[(\mathrm{E}_{\mathrm{T}_2} - \mathrm{E}_{\mathrm{T}_1})^2 + (\mathrm{V}_{\mathrm{T}_2} - \mathrm{V}_{\mathrm{T}_1})\left(\ln(\mathrm{V}_{\mathrm{T}_2}) - \ln(\mathrm{V}_{\mathrm{T}_1})\right)\right]^{1/2}
\end{aligned}
$$

and its associated probability of error $P_{e,lim}$, in the large n, p limit, thus reads

$$
P_{e,lim} = \frac{1}{2}\left(Q\left(\frac{s_{opt} - \mathrm{E}_{\mathrm{T}_1}}{\sqrt{\mathrm{V}_{\mathrm{T}_1}}}\right) + Q\left(\frac{\mathrm{E}_{\mathrm{T}_2} - s_{opt}}{\sqrt{\mathrm{V}_{\mathrm{T}_2}}}\right)\right). \qquad (14)
$$

Besides, recalling (12), and denoting $\mathbf{t}_V = p^{1/2}\,\mathbf{t}$, we have $\mathrm{E}_{\mathrm{T}_2} - \mathrm{E}_{\mathrm{T}_1} = \frac{\mathrm{f}''(\tau)}{\mathrm{p}^2}\,\gamma\,\boldsymbol{\ell}^\mathsf{T}\,\mathbf{P}_\mathrm{c}\,\mathbf{t}_{\Delta\mathrm{E}}$ with

$$
\begin{aligned}
\mathbf{t}_{\Delta\mathrm{E}} &= \mathrm{tr}\left(\mathbf{C}_{T_2} - \mathbf{C}_{T_1}\right)\,\mathbf{t}_V\\
&+ 2\,\left[\mathrm{tr}\left(\left(\mathbf{C}_{T_2} - \mathbf{C}_{T_1}\right)\mathbf{C}_1\right),\ \ldots,\ \mathrm{tr}\left(\left(\mathbf{C}_{T_2} - \mathbf{C}_{T_1}\right)\,\mathbf{C}_k\right)\right]^\mathsf{T}.
\end{aligned}
$$

Plugging in s_{opt} the quantities $\mathrm{V}_{\mathrm{T}_1}$ and $\mathrm{V}_{\mathrm{T}_2}$ by their respective expressions (13), $P_{e,lim}$ is then minimized for $\boldsymbol{\ell}_{opt}$ maximizing the Rayleigh quotient

$$
\boldsymbol{\ell}_{opt} = \arg\max_{\boldsymbol{\ell}} \frac{(\mathrm{E}_{\mathrm{T}_2} - \mathrm{E}_{\mathrm{T}_1})^2}{\mathrm{V}} = \frac{\boldsymbol{\ell}^\mathsf{T}\,\mathbf{P}_c\,\mathbf{t}_{\Delta\mathrm{E}}\,\mathbf{t}_{\Delta\mathrm{E}}^\mathsf{T}\,\mathbf{P}_c\,\boldsymbol{\ell}}{\boldsymbol{\ell}^\mathsf{T}\,\mathbf{P}_c\,\mathbf{t}_V\,\mathbf{t}_V^\mathsf{T}\,\mathbf{P}_c\,\boldsymbol{\ell}}. \qquad (15)
$$

As the matrix $\mathbf{t}_V\,\mathbf{t}_V^\mathsf{T}$ is of unit rank, the quotient is maximal for $\boldsymbol{\ell} = \boldsymbol{\ell}_{opt} \propto \left(\mathbf{t}_V\,\mathbf{t}_V^\mathsf{T}\right)^\dagger\,\mathbf{t}_{\Delta\mathrm{E}}$ with $(\cdot)^\dagger$ the pseudo-inverse.

4. CLASSIFICATION SIMULATED DATA AND POLSAR IMAGES

4.1. "Toy example" α-setting

Let us consider the following "toy-model" setup with target data randomly generated according to $\mathcal{N}(\mathbf{0}, \mathbf{C}_{T_1})$ and

[2] We can minimize the probability of error P_e by deriving $P_e(x) = 1/2\left(P\left(g(\mathbf{x}) > s \mid \mathbf{x} \in \mathcal{C}_{T_1}\right) + P\left(g(\mathbf{x}) < s \mid \mathbf{x} \in \mathcal{C}_{T_2}\right)\right)$. Function $g(\mathbf{x})$ having a Gaussian distribution, we have explicit access to $P_e(x)$ and we can derive it.

[3] The question of an optimal threshold based on class concentrations having already been studied in [5], we focus on the case of balanced classes.

$\mathcal{N}(\mathbf{0}, \mathbf{C}_{T_2})$ and source data drawn from $\mathcal{N}(\mathbf{0}, \mathbf{C}_{S_1})$ and $\mathcal{N}(\mathbf{0}, \mathbf{C}_{S_2})$. Target and source data are related through $\mathbf{C}_{T_1} = \alpha\,\mathbf{C}_{S_1} + (1-\alpha)\,\mathbf{C}_{S_2}$ and $\mathbf{C}_{T_2} = \alpha\,\mathbf{C}_{S_2} + (1-\alpha)\,\mathbf{C}_{S_1}$ with $\alpha \in [0, 1]$.

Figure 1 depicts the error rate empirically observed for $\boldsymbol{\ell} = [\ell_{S_1}, \ell_{T_1}, \ell_{S_2}, \ell_{T_2}]$, with $\boldsymbol{\ell} = [-1, -1, 1, 1]$, associating source S_a with target T_a, versus $\boldsymbol{\ell} = \boldsymbol{\ell}_{opt}$ obtained from (15). It is clearly observed that the optimized use of source data induces improved performance over the standard $\boldsymbol{\ell}_i \in \{\pm 1\}$ strategy, and over a target-only approach.

While the Gaussian mixture model might seem too restrictive, it has been recently proved, in the large dimensional setting under present concern, that most conventional machine learning algorithms treat advanced data models similarly to Gaussian mixture models: as such, these models are a sufficiently accurate assumption to absorb a large range of real data (starting with neural network features of real images, as proved by Seddik et al. in [11]).

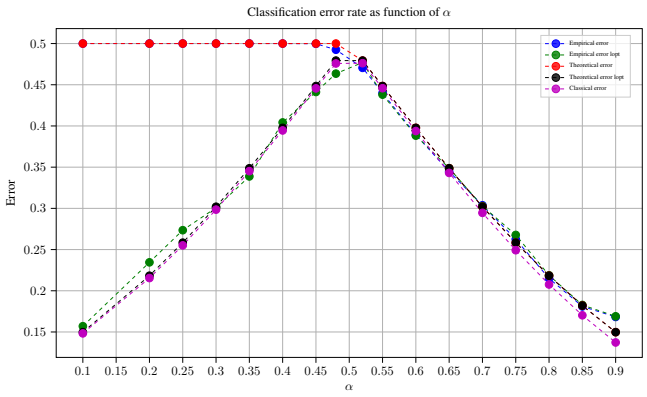

Fig. 1. Classification performance on simulated data, related by parameter α, s.t. $\mathbf{C}_{T_a} = \alpha\,\mathbf{C}_{S_a} + (1-\alpha)\,\mathbf{C}_{S_b}$, for various label strategies; $p = 512$, $n_{S_1} = n_{S_2} = 508$, $n_{T_1} = n_{T_2} = 4$, polynomial kernel f.

4.2. Experimental data

Our proposed transfer learning approach is here applied to a problem of terrain classification for polarimetric Synthetic Aperture Radar (PolSAR) images. A visual example is depicted at top of Figure 3, where the objective is to separate the two classes of terrain "road" and "fields".

The statistical (empirical) mean vectors in both classes are here very close. The complex-valued SAR data are characterized by their associated polarimetric channels of size $p_c = 3$. Using spatial (boxcar of size $p_{xy} = 3 \times 3$ pixels) diversities, the vector of real data characterizing each pixel has finally the size $p = 2\,p_c\,p_{xy} = 54$. The use of two adjacent spatial areas defines the two final data sets $\mathcal{C}_{T_1}, \mathcal{C}_{T_2}$ and $\mathcal{C}_{S_1}, \mathcal{C}_{S_2}$. According to the assumptions of 2.3, the data used here (source and target) are all centered (with zero mean vector) and we will use as kernel function $f(x) = (x - \tau)^2$, such that $f'(\tau) = 0$.

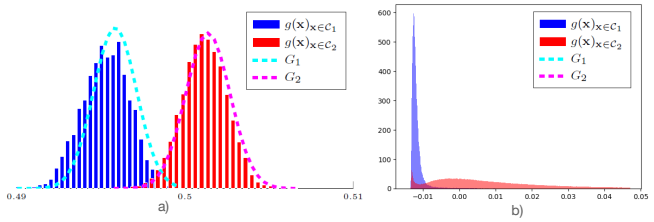

Fig. 2. Left: Histograms for Toy-example (Section 4.1). Right: Histograms from experimental SAR data (Section 4.2).

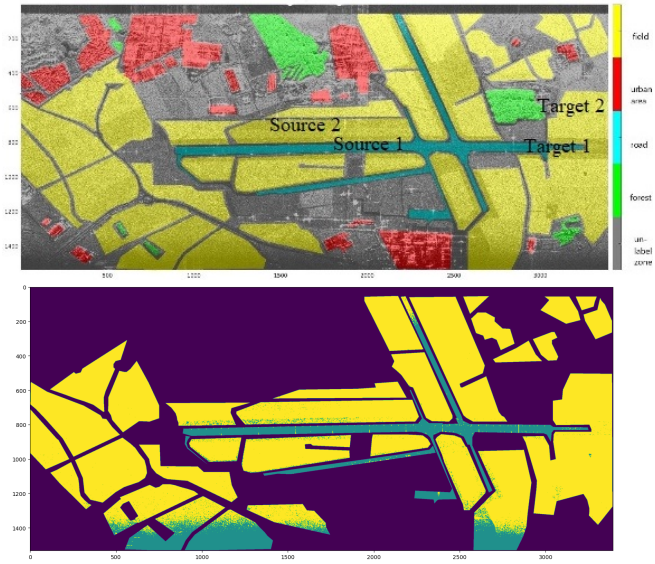

Fig. 3. Top: Classification between two target areas (Target 1 and Target 2) of a polarimetric SAR image of Bretigny provided by ONERA using source areas (Source 1 and Source 2). $p = 54$, $n = 2000$ (1000 per classes). Bottom: corresponding LS-SVM Classification

The separation of the two classes is illustrated through histograms on the right-hand side of Figure 2. The corresponding classification result is displayed at the bottom of Figure 3.

5. CONCLUDING REMARKS

This article proposed an improved LS-SVM transfer learning methodology for the problem of separating classes with equal statistical means. Our analysis revealed the importance to doctor the input data labels of both source and target data to maximize the classification performance. The analysis also emphasized the non-trivial impact of the "quality", largely privileged, over the quantity of the source data on the performance, which was verified on real PolSAR data. This process has been verified on simulated data and applied to experimental SAR images for POLSAR classification.

As such, the article provides new insights into transfer learning and the large degrees of achievable improvements, already in this quite elementary LS-SVM framework.

References

[1] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[3] Theodoros Evgeniou and Massimiliano Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 109–117.

[4] Romain Couillet, "A random matrix analysis and optimization framework to large dimensional transfer learning," in *IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 401–404.

[5] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1065–1074, 2019.

[6] Malik Tiomoko, Cosme Louart, and Romain Couillet, "Large dimensional asymptotics of multi-task learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8787–8791.

[7] Johan AK Suykens and Joos Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[8] Vladimir Vapnik, *The nature of statistical learning theory*, Springer science & business media, 1999.

[9] Romain Couillet and Florent Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[10] Tony Van Gestel, Johan AK Suykens, Gert Lanckriet, Annemie Lambrechts, Bart De Moor, and Joos Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis," *Neural computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[11] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet, "Random matrix theory proves that deep learning representations of gan-data behave as Gaussian mixtures," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8573–8582.