Apprentissage par transfert en grande dimension : application aux images SAR polarimétriques

Cyprien DOZ[1], Chengfang REN[1], Jean-Philippe OVARLEZ[1,2], Romain COUILLET[3]

[1]SONDRA, CentraleSupélec, Université Paris-Saclay

[2]DEMR, ONERA, Université Paris-Saclay

[3]GIPSA-Lab, Université Grenoble-Alpes

cyprien.doz@centralesupelec.fr, chengfang.ren@centralesupelec.fr
jeanphilippe.ovarlez@centralesupelec.fr, romain.couillet@gipsa-lab.grenoble-inp.fr

Résumé – Cet article présente une méthode d'apprentissage par transfert à noyau, sous un modèle de mélange gaussien à k classes pour les données d'entrée. En s'appuyant sur les avancées récentes de la théorie des matrices aléatoires, nous proposons de nouvelles perspectives dans les schémas d'apprentissage par transfert pour les cas difficiles, par exemple, ici, lorsque les statistiques de premier ordre de toutes les classes de données coïncident. L'article montre que la distribution asymptotique de la fonction de décision LS-SVM est gaussienne pour toute fonction à noyau. En conséquence, un schéma d'optimisation est proposé pour minimiser le taux d'erreur de classification. Nos résultats théoriques sont corroborés par des simulations, puis appliqués avec succès au contexte de l'apprentissage par transfert pour la classification d'images PolSAR.

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.

1 Introduction

En apprentissage automatique classique, les tâches sont généralement traitées séparément. Cette approche ne tient toutefois pas compte de la similarité potentiellement élevée entre les tâches. L'apprentissage par transfert vise à exploiter les informations contenues dans une tâche (source) pour aider à améliorer les performances de généralisation d'une autre tâche (*target*) ; voir [1] et [2] pour des tutoriels détaillés et diverses interprétations des méthodes actuelles.

Cet article se concentre spécifiquement sur une version *least-square support vector machine* (LS-SVM) de l'apprentissage par transfert, telle que présentée dans [3]. Cette dernière consiste à partager une partie de l'hyperplan de séparation dans les tâches sources et *target*, puis à résoudre une optimisation SVM parallèle pour les deux tâches, sous cette contrainte d'hyperplan partagé. Malgré la simplicité de la méthode, la compréhension du comportement et des performances de l'apprentissage par transfert reste très empirique. La question du choix approprié des hyperparamètres du SVM est particulièrement fondamentale afin d'éviter le redoutable problème du *transfert négatif*, par lequel l'optimisation de la tâche source entrave plutôt qu'elle n'aide la tâche *target*.

De premières percées ont été réalisées récemment dans [4, 5] en supposant un modèle statistique de grande dimension, en utilisant les avancées modernes en matière de matrice aléatoire. Le présent article exploite ces récentes découvertes pour généraliser le modèle LS-SVM à un mécanisme d'apprentissage par transfert *kernel LS-SVM*, naturellement approprié pour traiter des structures de données complexes. Plus précisément, l'article fournit une analyse théorique généralisant le modèle [6, 5] à un modèle de données avec différentes structures de covariance pour chaque classe de données.

Nos principales contributions peuvent être résumées comme suit : (i) dérivation de la performance asymptotique de l'apprentissage par transfert LS-SVM dans l'hypothèse de données nombreuses et de grande dimension ; (ii) analyse de la relation entre les structures des matrices de covariance dans toutes les classes de données (source et *target*) et le fonctionnement interne de l'apprentissage par transfert ; (iii) application à la classification des images SAR polarimétriques.

2 Modèle et hypothèses

2.1 Modèle des données

Soit $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un ensemble de n échantillons provenant de k classes, avec $\mathbf{x}_i \in \mathbb{R}^p$ et $y_i \in \{\ell_1, \ldots, \ell_k\}$ pour $\ell_a \in \mathbb{R}$ l'étiquette associée à la classe $\mathcal{C}_a$. Nous définissons $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}$ et $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_k]^\mathsf{T}$. On note $\mathbf{y} = \mathbf{J}\,\boldsymbol{\ell}$ avec, $\mathbf{J} \triangleq [\mathbf{j}_1, \ldots, \mathbf{j}_k]$, où $\mathbf{j}_a = \{\delta(y_i = \ell_a)\}_{i=1}^n$ est le vecteur indicateur[1] de la classe $\mathcal{C}_a$ de cardinalité n_a. On définit en outre $c_a = n_a/n$ pour $a \in \{1, \ldots, k\}$, $\mathbf{c} = [c_1, \ldots, c_k]^\mathsf{T}$,

1. $\delta(.)$ est la *delta fonction de Dirac*

$\mathbf{P} \triangleq \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ et $\mathbf{P}_c \triangleq \frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{P}\mathbf{J} = \mathrm{Diag}(\mathbf{c}) - \mathbf{c}\,\mathbf{c}^\mathsf{T}$.

L'ensemble de données est divisé entre un sous-ensemble *target*, qui contient des échantillons provenant exclusivement de deux classes et un sous-ensemble *source*, qui contient tous les autres échantillons, souvent nombreux, par rapport au sous-ensemble *target*.

Notre problème pratique est de classer de nouvelles données inconnues dans l'une des deux classes *target* tout en bénéficiant de l'existence des données sources étiquetées. La tâche qui nous intéresse est donc la classification dans les classes $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$, avec $T_1, T_2 \in \{1, \ldots, k\}$, tout en s'entraînant sur tous les échantillons disponibles, c'est-à-dire en incluant les données sources au processus d'entraînement. Afin de simplifier l'analyse, nous nous concentrons, sans perte de généralité, sur le cadre de $k = 4$ classes (dans la suite, nous désignerons ces indices de classe par S_1, T_1, S_2, T_2) : $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$ du côté de la cible et $\mathcal{C}_{S_1}$ et $\mathcal{C}_{S_2}$ du côté de la source. L'extension à un scénario multi-classes est cependant immédiate.

Les performances de classification sur $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$ vont naturellement dépendre fortement de la similarité de distribution des données dans les classes sources. Cette similarité est caractrisée par la matrice noyau $\mathbf{K}$, dont les entrées sont définies par $\mathbf{K}_{ij} = f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p\right)$, où $f : \mathbb{R} \to \mathbb{R}$ est la fonction noyau.

2.2 Décision par le LS-SVM

La méthode LS-SVM [7] a pour but de prédire l'étiquette de classe $\ell_\mathbf{x}$ des données entrantes $\mathbf{x}$, grâce à un apprentissage effectué sur le jeu de données d'entraînement $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, en concevant un hyperplan de séparation $\mathbf{w}^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}) + b$ entre $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$, qui est défini par le problème d'optimisation :

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

où $\gamma > 0$ est un hyperparamètre qui pénalise l'erreur de prédiction par rapport à la complexité potentielle du modèle. La résolution de (1) par le multiplicateur de Lagrange $\boldsymbol{\alpha}$ conduit à la solution $\hat{\mathbf{w}} = [\boldsymbol{\varphi}(\mathbf{x}_1)\ \cdots\ \boldsymbol{\varphi}(\mathbf{x}_n)]^\mathsf{T}\boldsymbol{\alpha} = \sum_{i=1}^n \alpha_i\boldsymbol{\varphi}(\mathbf{x}_i)$, où

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

avec $\mathbf{K} = \left\{\boldsymbol{\varphi}(\mathbf{x}_i)^\mathsf{T}\boldsymbol{\varphi}(\mathbf{x}_j)\right\}_{i,j}$ et $\mathbf{Q} = \left(\mathbf{K} + \frac{n}{\gamma}\mathbf{I}_n\right)^{-1}$. En utilisant le *kernel trick*, on remplace $\mathbf{K}$ par $\mathbf{K} = \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$ et $\mathbf{k}(\mathbf{x}) = \{f(\|\mathbf{x} - \mathbf{x}_j\|^2/p)\}_{j=1}^n$.

Étant donnés $\boldsymbol{\alpha}$ et b, une nouvelle donnée $\mathbf{x}$ est alors classée dans les classes $\mathcal{C}_a$ en fonction de la valeur de la fonction de décision

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

Pour classifier des données *target*, le seuil de décision de $g(\mathbf{x})$ est à définir que nous verrons dans la section 3 en s'appuyant de sa distribution asymptotique. En régime grande dimension, la plupart des méthodes d'apprentissage statistique, y compris le LS-SVM, considèrent que les algorithmes fonctionnent dans un régime $p \ll n$. Avant de passer aux résultats principaux de l'apprentissage par transfert LS-SVM, quelques hypothèses techniques supplémentaires pour se placer dans un régime de classification non trival.

2.3 Régime non trivial et hypothèses

Nous supposons que, pour $a \in \{1, \ldots, k\}$:

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

où $\boldsymbol{\mu} = \mathbf{0} \in \mathbb{R}^\mathbf{p}$ et $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, p^{-1}\,\mathbf{C}_a)$ avec $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ une matrice symétrique définie non-négative. Dans cet article, nous nous focaliserons donc sur le problème non trivial de séparation de données centrées, i.e. $\boldsymbol{\mu} = \mathbf{0}$.

Pour éviter des résultats triviaux, nous supposons, comme dans [8, 5], que les statistiques de classe satisfont certaines conditions de taux de croissance entre les matrices de covariances des données.

En posant $\mathbf{C}^\circ \triangleq \dfrac{1}{n}\sum_{a=1}^k n_a\,\mathbf{C}_a$ et en définissant le paramètre $\tau \triangleq \dfrac{2}{p}\,\mathrm{tr}\,\mathbf{C}^\circ > 0$, il a été montré dans [8] que, pour tout $i \neq j$:

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

Ce phénomène (problématique) de concentration de la distance permet d'approximer asymptotiquement $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$ par un développement de Taylor d'ordre deux de f autour de τ. On supposera ici que la fonction noyau f est au moins trois fois dérivable dans un voisinage de τ. Ainsi, le comportement asymptotique de l'apprentissage par transfert LS-SVM dépendra fortement des deux premières dérivées de f autour de τ.

Les performances de classification des LS-SVM en haute dimension dépendront fortement de la nature de la fonction noyau. Cette caractéristique devient adaptable à différents problèmes donnés comme exploré dans [8] et [5] pour une différence évanescente des moyennes entre les classes.

Dans un cadre statistique classique, les caractéristiques du cadre SVM ont fait de l'analyse de ses performances une tâche difficile. Les performances des SVM ont été étudiées avec différentes approches (en introduisant la dimension VC [9] ou l'interprétation bayésienne [10]) en gardant p et n fixes. L'analyse asymptotique récente menée dans [5] a ouvert la porte à des pistes d'amélioration pour le cadre LS-SVM. Maintenant, nous avons l'intention de mener le même type d'analyse asymptotique afin d'explorer le fonctionnement interne entre les données source et cible dans un contexte d'apprentissage par transfert. Les performances de classification des LS-SVM en haute dimension dépendent fortement de la nature de la fonction noyau. Cette caractéristique devient adaptable à différents problèmes donnés comme exploré dans [8]. Avec les hy-

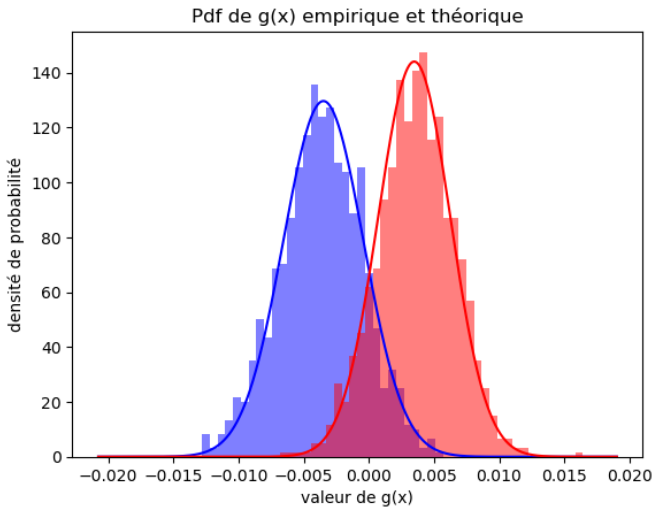

FIGURE 1 – Evaluation des performances par $\mathcal{N}(E_a, V_a)$ pour les classes $\mathcal{C}_{T_1}$ (en bleu) et $\mathcal{C}_{T_2}$ (en rouge).

pothèses précédentes, nous sommes maintenant capables d'effectuer une analyse technique de notre problème d'apprentissage par transfert LS-SVM en haute dimension.

3 Résultats asymptotiques

Afin d'évaluer les performances du classifieur et l'impact des données sources dans l'ensemble d'apprentissage, lorsque $n, p \to \infty$, notre objectif est de fournir une approximation de $g(\mathbf{x})$ dans ce régime. La solution $(\boldsymbol{\alpha}, b)$ de (2) est une fonction de $\mathbf{y}$ et de $\mathbf{Q}$. Comme $\mathbf{Q}$ n'est pas directement accessible, nous procédons en deux étapes : (i) en exploitant les résultats de [8], nous exploitons une "linéarisation" asymptotique techniquement commode de $\mathbf{K}$, (ii) qui permet ensuite d'effectuer une expansion de Taylor sur $\mathbf{Q}$ autour de son terme dominant $\left(f(\tau)\,\mathbf{1}_n\,\mathbf{1}_n^\mathsf{T} + \frac{n}{\gamma}\,\mathbf{I}_n\right)^{-1}$. En procédant comme dans [5], nous obtenons ensuite des approximations précises, dans un cadre asymptotique, pour $\boldsymbol{\alpha}$ et b. Dans le cas d'applications du papier, nous nous concentrons sur la classification de données à vecteur moyenne nul (et donc seulement discriminées par les covariances de classe). Conformément aux remarques faites notamment dans [8] et [5]. nous avons démontré qu'il était plus intéressant d'utiliser des noyaux vérifiant $f'(\tau) = 0$. En posant :

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

le résultat principal de [5] s'étend comme suit :

Théorème 1 (Approximation Gaussienne) *Soit* $\mathbf{x} \in \mathcal{C}_a, a \in \{T_1, T_2\}$. *En considérant vérifiées les hypothèses enoncées, la loi asymptotique de* $g(\mathbf{x})$ *défini par* (3) *est donnée par :*

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

où la moyenne E_a *et la variance* V_a *sont définis comme*

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

Ce théorème nous permet de pouvoir caractériser les performances de classification pour les deux classes (voir Fig. 1). On peut ainsi également caractériser le seuil de décision optimal qui permet de séparer les classes $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$. Si on désigne par s le seuil de décision d'appartenance à la classe, l'erreur de classification P_e est donnée par

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

Avec le résultat du Theorème 1 et après quelques manipulations, on peut définir le seuil optimal s_{opt} de classification entre les deux variables gaussiennes $G_1 \sim \mathcal{N}(\mathrm{E}_{T_1}, \mathrm{V}_{T_1})$ et $G_2 \sim \mathcal{N}(\mathrm{E}_{T_2}, \mathrm{V}_{T_2})$ qui minimise la probabilité P_e :

$$
s_{opt} = \frac{\mathrm{E}_{T_2}\mathrm{V}_{T_1} - \mathrm{E}_{T_1}\mathrm{V}_{T_2}}{\mathrm{V}_{T_1} - \mathrm{V}_{T_2}} - \frac{(\mathrm{V}_{T_1}\mathrm{V}_{T_2})^{1/2}}{\mathrm{V}_{T_1} - \mathrm{V}_{T_2}} \times
$$
$$
\left[(\mathrm{E}_{T_2} - \mathrm{E}_{T_1})^2 + (\mathrm{V}_{T_2} - \mathrm{V}_{T_1})\left(\ln(\mathrm{V}_{T_2}) - \ln(\mathrm{V}_{T_1})\right)\right]^{1/2}
$$

Dans ce régime spécifique, la probabilité d'erreur asymptotique va dépendre des labels. Il est ainsi possible d'établir des stratégies optimales d'étiquetage des données après avoir étudié leurs caractéristiques les unes par rapport aux autres.

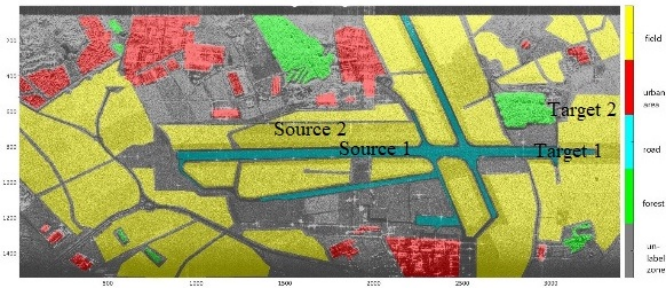

FIGURE 2 – Classification entre deux zones cibles (Target1 et Target2) d'une image SAR polarimétrique de Bretigny fournie par l'ONERA à l'aide zones sources (Source1 et Source2). $p = 54$, $n = 2000$ (1000 par classe).

4 Classification d'images PolSAR

L'approche d'apprentissage par transfert que nous proposons est ici appliquée à un problème de classification de terrain pour des images polarimétriques de radar à ouverture synthétique (PolSAR). Un exemple visuel est représenté sur la Figure 2 où l'objectif est de séparer les deux classes de terrain présentées. Les vecteurs moyens statistiques (empiriques) des deux classes sont ici très proches. Les données SAR étant complexes (deux canaux) et polarimétriques (trois canaux), la taille des vecteurs

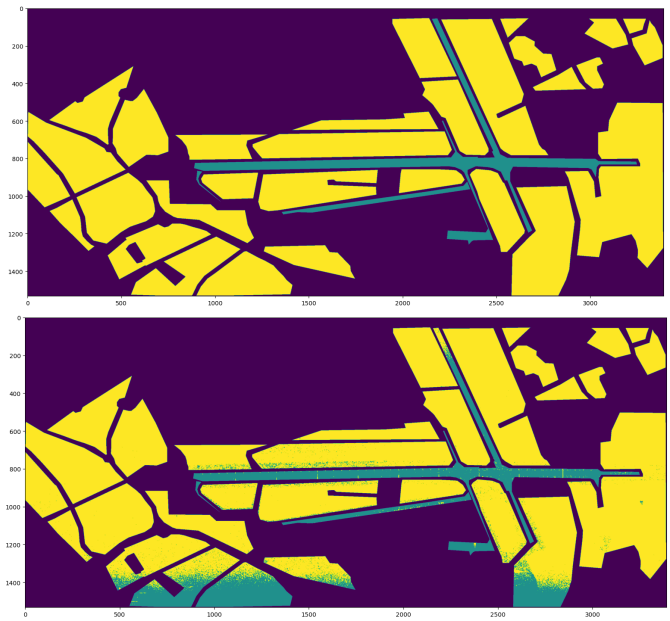

FIGURE 3 – Haut : Vérité terrain. Bas : Classification LS-SVM

de données réelles est $p_c = 6$. En concaténant ces données dans une petite fenêtre spatiale carrée de taille $p_{xy} = 3 \times 3$ pixels), le vecteur caractérisant chaque pixel a finalement la taille $p = p_c\, p_{xy} = 54$.

L'utilisation de deux zones spatiales adjacentes définit les deux ensembles finaux de données $\mathcal{C}_{T_1}, \mathcal{C}_{T_2}$ et $\mathcal{C}_{S_1}, \mathcal{C}_{S_2}$. Conformément aux hypothèses de 2.3, les données utilisées (source et target) sont ici toutes centrées (à vecteur moyenne nulle) et on utilisera comme fonction noyau $f(x) = (x - \tau)^2$, telle que $f'(\tau) = 0$. Les valeurs de la fonction de décision sont affichées en Fig. 4. On constate que sur des données réelles, les valeurs se répartissent également selon deux gaussiennes. Cela permet d'avoir accès à la probabilité d'erreur de classification ainsi que le seuil de classification optimal.

5 Conclusion

Cet article propose une méthodologie améliorée d'apprentissage par transfert LS-SVM pour séparer des classes ayant des moyennes statistiques égales. Notre analyse souligne également l'importance d'une analyse asymptotique pour anticiper les performances de l'algorithme. Nous avons pu illustré la pertinence de l'approche sur des données PolSAR réelles.

Notre travail propose par conséquent de nouvelles perspectives dans le domaine de l'apprentissage par transfert, et il ouvre la voie à des améliorations réalisables dans ce cadre élémentaire de LS-SVM.

Références

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

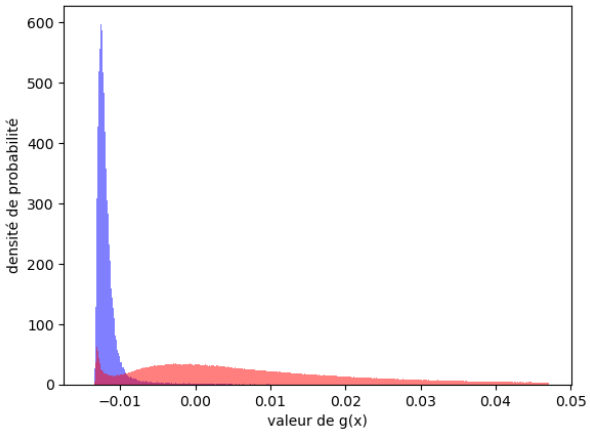

FIGURE 4 – Densités de probabilité des classes $\mathcal{C}_{T_1}$ et $\mathcal{C}_{T_2}$

[2] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

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

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

[5] Z. Liao and R. 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] M. Tiomoko, C. Louart, and R. Couillet, "Large dimensional asymptotics of multi-task learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8787–8791, IEEE, 2020.

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

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

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

[10] T. Van Gestel, J. A. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. 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.