Étude de la généralisation de DASF à l'adaptation de domaine semi-supervisée

Emilie Morvant, Amaury Habrard, Stéphane Ayache

Étude de la généralisation de DASF à l'adaptation de domaine semi-supervisée

Emilie Morvant[1], Amaury Habrard[2], Stéphane Ayache[1]

[1] Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France

[2] Université de St-Etienne, Lab. Hubert Curien, CNRS, UMR 5516, F-42000, St-Etienne, France

{prenom.nom}@lif.univ-mrs.fr, amaury.habrard@univ-st-etienne.fr

Résumé : L'adaptation de domaine est un problème important en apprentissage automatique. Il s'intéresse au cas où les données de test sont générées selon une distribution (cible) différente de celle ayant généré les données d'apprentissage (source). Dans ce cadre, Ben-David *et al.* ont montré qu'un classifieur a de meilleures garanties de généralisation lorsque la distance entre les deux distributions marginales selon l'espace d'entrée est faible. Dans le cas non-supervisé, lorsque les données d'apprentissage étiquetées sont uniquement issues de la distribution source, Morvant *et al.* ont créé un algorithme itératif - DASF - visant à diminuer cette distance par la construction d'un espace de projection explicite défini par une bonne fonction de similarité (au sens de Balcan *et al.*). Dans cet article nous généralisons DASF au cas semi-supervisé dans lequel quelques données cibles d'apprentissage sont étiquetées. Notre méthode se base sur le cadre théorique de Ben-David *et al.* proposant la minimisation d'une combinaison convexe des erreurs empiriques source et cible. Nous réalisons une étude de la parcimonie et de la capacité en généralisation des modèles inférés par notre méthode, puis nous confirmons cette analyse sur un exemple jouet et une tâche d'annotation réelle. **Mots-clés** : Apprentissage par Transfert, Adaptation de Domaine, Fonction de Similarité.

1. Introduction

Pour classer automatiquement des données, une solution est d'apprendre un classifieur à partir d'exemples étiquetés et représentatifs des données à classer. En d'autres termes, les données d'apprentissage et de test suivent la même distribution. Cependant, étiqueter des exemples pertinents reste en pratique coûteux. Dans une telle situation, l'*apprentissage par transfert* (Pan & Yang, 2010), plus précisément l'*adaptation de domaine* (AD), propose d'adapter un modèle d'une distribution d'apprentissage – le *domaine source* –

vers une distribution test différente – le *domaine cible* (Jiang, 2008; Quionero-Candela *et al.*, 2009). On distingue deux cas : si les données d'apprentissage étiquetées sont issues du domaine source, on parle d'*AD non-supervisée* (non-sup.) ; si des données étiquetées cibles sont disponibles, on parle d'*AD semi-supervisée* (semi-sup.). D'un point de vue général, lorsque les distributions marginales selon l'espace d'entrée sont proches et qu'il existe un classifieur performant sur le domaine source, alors Ben-David *et al.* (2010a) et Mansour *et al.* (2009) ont démontré que les garanties de ce classifieur sur le domaine cible peuvent être bonnes. Ainsi, une solution à l'AD consiste à rapprocher les distributions tout en gardant de bonnes performances sur le domaine source. Pour le cas de l'AD non-sup., différentes méthodes ont été proposées. Une idée consiste à repondérer les données sources pour les rapprocher des cibles (Mansour *et al.*, 2009; Huang *et al.*, 2006; Sugiyama *et al.*, 2007; Jiang & Zhai, 2007), une autre construit un nouvel espace où les distributions marginales sont proches (Ben-David *et al.*, 2010a; Blitzer *et al.*, 2011). C'est le cas de l'algorithme DASF *(Domain Adaptation with Similarity Function)* (Morvant *et al.*, 2011, 2012) qui exploite l'apprentissage avec une *bonne fonction de similarité* (Balcan *et al.*, 2008), laquelle, associée à des points *landmarks*, définit un espace de projection explicite où il existe un classifieur performant. Morvant *et al.* (2011, 2012) proposent donc de modifier cet espace pour rapprocher les distributions à l'aide d'un terme de régularisation se focalisant sur les landmarks à la fois similaires à des points sources et à des points cibles.

Pour améliorer la recherche du classifieur, les étiquettes cibles apportent naturellement une information cruciale. Un bon algorithme d'AD doit donc être capable de considérer le cas semi-sup. Dans ce but, des méthodes étendent linéairement l'espace de projection (Daumé III, 2007; Daumé III *et al.*, 2010), d'autres combinent de l'information d'étiquettes sources et cibles (Bergamo & Torresani, 2010; Daumé III *et al.*, 2010). Ben-David *et al.* (2010a) proposent quant à eux de minimiser une combinaison convexe des erreurs empiriques source et cible. En suivant ce cadre, nous étendons DASF au cas semi-sup. De plus, nous réalisons une analyse théorique puis empirique de la capacité en généralisation de notre méthode et de la parcimonie des modèles inférés. Nous fondons cette étude sur le cadre de la *robustesse* (Xu & Mannor, 2012).

Tout d'abord, nous énonçons les cadres d'AD de Ben-David *et al.* en Sec.2. et des fonctions de similarité de Balcan *et al.* en Sec.3. Ensuite nous présentons successivement, en Sec.4., l'algorithme non-sup. DASF de Morvant *et al.* puis notre généralisation semi-sup. En Sec.5., nous en analysons théoriquement le comportement et complétons empiriquement cette étude en Sec.6.

2. Adaptation de domaine

Soit X l'espace d'entrée et $Y=\{-1,1\}$ les étiquettes. Un domaine est une distribution de probabilité sur $X\times Y$. En AD, les domaines différents P_S et P_T désignent respectivement le *domaine source* et le *domaine cible*. D_S et D_T sont les distributions marginales associées selon X. En AD semi-sup., quelques étiquettes cibles sont disponibles. Ainsi, on pose un *échantillon étiqueté* $LS=(LS_S,LS_T)$ tel que $LS_S=\left\{(\mathbf{x}_{i^S},y_{i^S})\right\}_{i^S=1}^{d_l^S}$ et $LS_T=\left\{(\mathbf{x}_{i^T},y_{i^T})\right\}_{i^T=1}^{d_l^T}$ contiennent respectivement d_l^S données sources *i.i.d.* selon P_S et d_l^T données cibles *i.i.d.* selon P_T, ainsi qu'un *échantillon non-étiqueté cible* $TS=\{\mathbf{x}_{i'}\}_{i'=1}^{d_t}$ *i.i.d.* selon D_T. On suppose $d_l^T \ll d_l^S$, sinon un algorithme usuel d'apprentissage supervisé sera préféré. Notons que $d_l^T = 0$ revient au cas non-sup. Soit une hypothèse $h : X \to Y$, ici un classifieur binaire. Les erreurs réelles source et cible de h sont la probabilité que h commette une erreur sur respectivement P_S et P_T : $err_S(h)=\mathbb{E}_{(\mathbf{x},y)\sim P_S} L_{01}(h,(\mathbf{x},y))$, $err_T(h)=\mathbb{E}_{(\mathbf{x},y)\sim P_T} L_{01}(h,(\mathbf{x},y))$, où $L_{01}(h,(\mathbf{x},y))$, la *fonction perte* 0-1, vaut 1 si $h(\mathbf{x}) \neq y$, 0 sinon. Les erreurs empiriques associées sont $\hat{err}_S$ et $\hat{err}_T$. Soit $\mathcal{H}$ une classe d'hypothèses de X vers Y, le but de l'AD est donc de trouver $h \in \mathcal{H}$ minimisant l'erreur cible $err_T(h)$. Nous considérons la théorie générale d'AD de Ben-David *et al.* (2010a) dont le résultat principal est la borne suivante, majorant l'erreur cible en fonction de l'erreur source $err_S(h)$ et d'une divergence entre les domaines,

$$\forall h \in \mathcal{H},\ err_T(h) \leq err_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu. \qquad (1)$$

La contante ν, liée à la capacité d'adaptation de $\mathcal{H}$ pour le problème considéré, est l'erreur de l'hypothèse jointe idéale sur les deux domaines : $\nu = err_S(h^*) + err_T(h^*)$, $h^* = \text{argmin}_{h\in\mathcal{H}}(err_S(h) + err_T(h))$. Le terme $d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T)$, dépendant de $\mathcal{H}$, est la $\mathcal{H}\Delta\mathcal{H}$-distance entre les distributions marginales. Si $\mathcal{H}$ est de dimension VC finie, elle peut s'estimer en cherchant le meilleur classifieur séparant un échantillon source fini d'un échantillon cible fini. Finalement, un principe, nécessaire pour l'AD (Ben-David *et al.*, 2010b), vise à inférer un espace tel qu'à la fois la $\mathcal{H}\Delta\mathcal{H}$-distance et l'erreur source soient faibles.
Dans le contexte particulier de l'AD semi-sup., l'information des étiquettes cibles peut aider à la recherche du classifieur. Or, minimiser uniquement l'erreur empirique cible depuis LS_T n'est pas la meilleure solution car LS_T n'est pas représentatif de P_T. Ben-David *et al.* (2010a) proposent donc de considérer une combinaison convexe des erreurs empiriques source et cible,

$$\kappa \in [0,1],\ \forall h \in \mathcal{H},\ \hat{err}_\kappa(h) = \kappa\, \hat{err}_T(h) + (1-\kappa)\, \hat{err}_S(h), \qquad (2)$$

On note $err_\kappa(h)$ l'erreur réelle pondérée associée. La borne suivante relie $err_T(h)$ et $err_\kappa(h)$, justifiant de la cohérence de la minimisation de l'Eq.(2).

$$\forall h \in \mathcal{H},\ |err_\kappa(h) - err_T(h)| \leq (1-\kappa)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu\right). \quad (3)$$

Lorsque κ tend vers 1, les données cibles sont de plus en plus importantes, alors que les données sources et la $\mathcal{H}\Delta\mathcal{H}$-distance le sont de moins en moins.

Nous présenterons en Sec.4.2. notre méthode permettant de généraliser DASF de Morvant *et al.* (2011) à ce cadre semi-sup.

3. Apprentissage supervisé avec une bonne fonction de similarité

La méthode DASF présentée en Sec.4.1. est basée sur un cadre permettant d'éviter les contraintes liées à l'espace de projection implicite et de grande dimension associé aux noyaux usuels (symétrie et semi-definie positivité (SDP)). Ce cadre proposé par Balcan *et al.* (2008) introduit une notion plus intuitive et plus flexible de *bonne fonction de similarité* selon la définition suivante :

Définition 1 (Balcan *et al.* (2008))
Une fonction de similarité sur X est une fonction $K : X \times X \to [-1,1]$. K est dites **(ϵ,γ,τ)-bonne** *sur un domaine P, s'il existe une fonction indicatrice aléatoire $R(\mathbf{x})$ définissant un ensemble de points raisonnables tel que :*
(i) un taux de $1-\epsilon$ des points $(\mathbf{x},y)$ vérifient, $\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[yy'K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')\big] \geq \gamma$,
(ii) $Pr_{\mathbf{x}'}[R(\mathbf{x}')] \geq \tau$.

En d'autres termes : *(i)* avec une confiance γ, la majorité des exemples sont plus similaires aux points raisonnables de même classe qu'à ceux de classe opposée ; *(ii)* au moins une proportion τ des points sont raisonnables. La Def.1 inclut aussi bien les noyaux valides que des similarités non-SDP et non-symétriques et généralise ainsi les noyaux. En pratique, les points raisonnables sont inconnus, on notera donc $R=\{\mathbf{x}'_j\}_{j=1}^{d_u}$ un ensemble aléatoire de d_u points potentiellement raisonnables dîts *landmarks*. Si K est (ϵ,γ,τ)-bonne, alors *(i)* et *(ii)* sont suffisantes pour apprendre avec une grande probabilité un séparateur performant dans un ϕ^R-espace : $\phi^R : \begin{cases} X \to \mathbb{R}^{d_u} \\ \mathbf{x} \mapsto \langle K(\mathbf{x},\mathbf{x}'_1),..,K(\mathbf{x},\mathbf{x}'_{d_u})\rangle \end{cases}$.
Soit $LS=\big\{(\mathbf{x}_i,y_i)\big\}_{i=1}^{d_l}\sim P$, un tel séparateur $\boldsymbol{\alpha}\in\mathbb{R}^{d_u}$ s'apprend alors en résolvant un problème linéaire (Balcan *et al.*, 2008), dont une formulation est,

$$\left\{\begin{array}{l} \min\limits_{\boldsymbol{\alpha}} \dfrac{1}{d_l}\sum\limits_{i=1}^{d_l} L\big(g,(\mathbf{x}_i,y_i)\big) + \lambda\|\boldsymbol{\alpha}\|_1, \\ \text{avec } L\big(g,(\mathbf{x}_i,y_i)\big) = \Big[1-y_i g(\mathbf{x}_i)\Big]_+ \text{ et } g(\mathbf{x}_i) = \sum\limits_{j=1}^{d_u}\alpha_j K(\mathbf{x}_i,\mathbf{x}'_j), \end{array}\right. \quad (SF_{opt})$$

où g est le modèle inféré. Par la suite, nous notons "classifieur-SF" un classifieur $h(\mathbf{x}) = \text{sign}[g(\mathbf{x})]$ appris par (SF_{opt}) et $\mathcal{H}$ la classe des classifieurs-SF. Remarquons qu'en considérant le problème d'AD visant à la minimisation de la borne (1)., (SF_{opt}) peut être vu comme la minimisation empirique de l'erreur sur le domaine source $P_S = P$, tout en définissant un ϕ^R-espace pertinent puisque les landmarks de poids nul dans la solution $\boldsymbol{\alpha}$ ne sont pas considérés. Une idée est donc de contraindre le ϕ^R-espace à diminuer la $\mathcal{H}\Delta\mathcal{H}$-distance.

4. Adaptation de domaine avec une bonne fonction de similarité

Tout d'abord, nous présentons la méthode DASF (Morvant *et al.*, 2011) basée sur un terme de régularisation sur $\boldsymbol{\alpha}$ contraignant le ϕ^R-espace à abaisser la $\mathcal{H}\Delta\mathcal{H}$-distance. Ensuite, nous proposons une généralisation de DASF au cas semi-sup. fondée sur le résultat de l'Eq.(3) de Ben-David *et al.* (2010a).

4.1. DASF : Adaptation de domaine non-supervisée

En AD non-sup., on a $d_l^T = 0$ et donc $LS = LS_S = \big\{(\mathbf{x}_i,y_i)\big\}_{i=1}^{d_l} \sim P_S$ est l'échantillon source de d_l exemples étiquetés ($LS_{S|X} = \big\{\mathbf{x}_i/(\mathbf{x}_i,y_i)\in LS_S\big\}_{i=1}^{d_l}$). Nous posons $R = \big\{\mathbf{x}'_j\big\}_{j=1}^{d_u}$ les landmarks et $TS \sim D_T$ l'échantillon non-étiqueté cible. Le régularisateur additionel est construit pour rapprocher $LS_{S|X}$ et TS. Morvant *et al.* (2011) ont ainsi exploité la notion de *robustesse d'algorithme* (Xu & Mannor, 2012) : *"If a testing sample is similar to a training sample then the testing error is close to the training error."* (*cf.* Def.2). Ce principe suggère que $LS_{S|X}$ et TS peuvent être similaires si pour des paires de points source et cible proches et de même classe $(\mathbf{x}_s,\mathbf{x}_t)$, l'écart entre les pertes de $\mathbf{x}_s$ et $\mathbf{x}_t$ est faible. En considérant la perte hinge de (SF_{opt}), pour tout modèle appris g et toutes paires $(\mathbf{x}_s,\mathbf{x}_t)$ de classe y, le terme de régularisation est basé sur une majoration de cet écart entre les pertes : $|L(g,(\mathbf{x}_s,y)) - L(g,(\mathbf{x}_t,y))| \leq \|({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t))\,\text{diag}(\boldsymbol{\alpha})\|_1$, où ${}^t\phi^R(\cdot)$ est le vecteur transposé de $\phi^R(\cdot)$ et $\text{diag}(\boldsymbol{\alpha})$ la matrice diagonale de coefficients $\boldsymbol{\alpha}$. Étant donné un ensemble de paires $\mathcal{C}_{ST} \subset LS_{S|X} \times TS$, le problème d'optimisation d'AD non-sup. est obtenu en ajoutant, à (SF_{opt}), le régularisateur pour chaque paire.

$$\begin{cases} \min_{\boldsymbol{\alpha}} \frac{1}{d_l} \sum_{i=1}^{d_l} L\big(g,(\mathbf{x}_i,y_i)\big) + \lambda\|\boldsymbol{\alpha}\|_1 + \beta \sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \Big\| \big({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t)\big)\mathrm{diag}(\boldsymbol{\alpha}) \Big\|_1, \\ \text{avec } L\big(g,(\mathbf{x}_i,y_i)\big) = \Big[1 - y_i g(\mathbf{x}_i)\Big]_+ \text{ et } g(\mathbf{x}_i) = \sum_{j=1}^{d_u} \alpha_j K(\mathbf{x}_i,\mathbf{x}'_j). \end{cases} \quad (\textit{DASF}_{opt})$$

La résolution de (*DASF*$_{opt}$) construit un ϕ^R-espace en essayant de minimiser deux termes de la borne d'AD (1) – l'erreur source err_S et la $\mathcal{H}\Delta\mathcal{H}$-distance – en rapprochant les points de $\mathcal{C}_{ST}$. Les paires de $\mathcal{C}_{ST}$ sont choisies parmi des points proches dans les échantillons source et cible. En pratique le nombre de paires possibles est trop important et seulement un nombre limité de points sera considéré. La possible perte d'information est alors compensée par un processus itératif pour compléter la recherche d'un espace pertinent. Ce processus repose sur le fait qu'à une itération l donnée, le terme de régularisation proposé peut être vu comme la minimisation d'une distance $L1$ dans un nouvel ϕ^R_{l+1}-espace : $\|({}^t\phi^R_l(\mathbf{x}_s) - \phi^R_l(\mathbf{x}_t))\,\mathrm{diag}(\boldsymbol{\alpha}^l)\|_1 = \|{}^t\phi^R_{l+1}(\mathbf{x}_s) - \phi^R_{l+1}(\mathbf{x}_t)\|_1$, ϕ^R_{l+1} étant défini par la similarité K_{l+1} obtenue en pondérant K_l conditionnellement à chaque landmark : $\forall \mathbf{x}'_j \in R$, $K_{l+1}(\cdot,\mathbf{x}'_j) = \alpha^l_j K_l(\cdot,\mathbf{x}'_j)$. On itère ensuite dans le nouvel ϕ^R_{l+1}-espace. Les hyperparamètres sont estimés à l'aide d'un principe de validation inverse basé sur l'erreur que ferait un classifieur appris uniquement sur des exemples cibles étiquetés par le modèle trouvé par (*DASF*$_{opt}$). L'algorithme utilise une estimation heuristique de l'erreur optimale jointe ν de la borne d'AD (1) et choisit les paramètres minimisant ce critère, le principe itératif s'arrête lorsque l'estimation de ν ne décroit plus.

4.2. SSDASF : Adaptation de domaine semi-supervisée

D'après le cadre de l'AD semi-sup. proposé par Ben-David *et al.* (2010a), nous modifions (*DASF*$_{opt}$) permettant l'utilisation d'étiquettes cibles. Étant donné un échantillon étiqueté $LS{=}(LS_S,LS_T)$ de $d_l = d_l^S + d_l^T$ points, un ensemble de paires $\mathcal{C}_{ST} \subset LS_{S|X} \times TS$ et $\kappa \in [0,1]$, nous définissons le problème suivant.

$$\begin{cases} \min_{\boldsymbol{\alpha}} (1-\kappa)\frac{1}{d_l^S} \sum_{i^S=1}^{d_l^S} L\big(g,(\mathbf{x}_{i^S},y_{i^S})\big) + \kappa\frac{1}{d_l^T} \sum_{i^T=1}^{d_l^T} L\big(g,(\mathbf{x}_{i^T},y_{i^T})\big) + \lambda\|\boldsymbol{\alpha}\|_1 \\ \qquad +(1-\kappa)\beta \sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \Big\| \big({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t)\big)\,\mathrm{diag}(\boldsymbol{\alpha}) \Big\|_1, \\ \text{avec } L\big(g,(\mathbf{x}_i,y_i)\big) = \Big[1 - y_i g(\mathbf{x}_i)\Big]_+ \text{ et } g(\mathbf{x}_i) = \sum_{j=1}^{d_u} \alpha_j K(\mathbf{x}_i,\mathbf{x}'_j). \end{cases} \quad (\textit{SSDASF}_{opt})$$

Alors que (*DASF*$_{opt}$) se focalise uniquement sur la minimisation de l'erreur empirique source, (*SSDASF*$_{opt}$) minimise la combinaison convexe des erreurs empiriques source et cible (*cf.* Eq.(2)). Il existe un lien entre les deux problèmes. En effet, si $\kappa = 0$, aucune étiquette cible n'est utilisée, nous revenons à (*DASF*$_{opt}$). Dans le cas contraire, si $\kappa = 1$, nous arrivons à un cadre d'apprentissage supervisé usuel pour lequel les échantillons d'apprentissage et de test sont tirés selon le même domaine P_T. L'Algo.1 décrit l'algorithme de notre méthode semi-sup. (SSDASF). Les principes itératif et de validation de DASF sont suivis, seul le problème d'optimisation au cœur de la méthode est modifié.

Algorithme 1 SSDASF : Semi-Supervised Domain Adaptation with Similarity Function

entrée Fonction de similarité K, ensembles R, $LS = (LS_S, LS_T)$ et TS

$h_0(\cdot) \leftarrow \text{sign}\left[\frac{1}{|R|}\sum_{j=1}^{|R|} K(\cdot, \mathbf{x}'_j)\right]$; $K_1 \leftarrow K$; $l \leftarrow 1$

tant que Le critère d'arrêt n'est pas vérifié **faire**

Construction de $\mathcal{C}_{ST}$

$\boldsymbol{\alpha}^l \leftarrow$ Résoudre (*SSDASF*$_{opt}$) avec K_l et $\mathcal{C}_{ST}$

$K_{l+1} \leftarrow$ MAJ de K_l en fonction de $\boldsymbol{\alpha}^l$; MAJ de R ; $l++$

fin tant que

retourner $h(\cdot) = \text{sign}\left[\sum_{x'_j \in R} \alpha_j^l K_l(\cdot, \mathbf{x}'_j)\right]$

5. Étude théorique

Nous proposons dans cette section une étude de la parcimonie et de la capacité en généralisation du problème (*SSDASF*$_{opt}$). Pour espérer une bonne AD, $\mathcal{C}_{ST}$ doit être informatif. Nous supposons donc que pour chaque coordonnée $\mathbf{x}'_j$ du ϕ^R-espace, il existe au moins une paire avec de différentes valeurs : $\forall \mathbf{x}'_j \in R$, $\max_{(\mathbf{x}_s \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| > 0$. En AD, les domaines étant *a priori* différents, cette hypothèse arrive donc avec une grande probabilité.

Analyse de la parcimonie. (*SSDASF*$_{opt}$) comporte deux termes de régularisation de norme 1 contenant $\boldsymbol{\alpha}$, le premier, $\|\boldsymbol{\alpha}\|_1$, implique une parcimonie naturelle. Le lemme suivant nous permet de considérer l'influence des deux termes.

Lemme 1

*Pour tous les hyperparamètres $\lambda > 0$, $\beta > 0$ et $\kappa \in [0, 1]$ et pour tout ensemble de paires $\mathcal{C}_{ST}$, on pose $B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| \right\}$. Si $\boldsymbol{\alpha}^*$ est la solution optimale de (*SSDASF*$_{opt}$), alors on a, $\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{(1-\kappa)\beta B_R + \lambda}$.*

Démonstration. Voir Morvant *et al.* (2012). ■

Ce lemme montre que la parcimonie du modèle dépend des hyperparamètres λ, β, κ et de la quantité B_R, quantité liée à la distance entre les points des paires de $\mathcal{C}_{ST}$ dans le ϕ^R-espace (B_R est le minimum des déviations maximales des coordonnées des points d'une même paire). Plus les distributions marginales sont éloignées, plus la tâche est dure et plus B_R tend à croître, impliquant une parcimonie plus forte. Si $\kappa = 0$, alors nous sommes dans le cas non-sup. de ($DASF_{opt}$) et la borne devient $\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}$. Si $\kappa > 0$, alors la prise en compte de données cibles implique un modèle moins parcimonieux.

Bornes en généralisation. Tout d'abord, nous rappelons la définition d'un *algorithme (pseudo-)robuste* et le théorème sur la capacité en généralisation d'un tel algorithme dans le contexte usuel où les données d'apprentissage et de test sont issues d'un même domaine (Xu & Mannor, 2012). Ce cadre révèle en fait deux avantages : la prise en compte des termes de régularisation dans la borne et l'étude de contextes non-standards tel que l'AD.

Définition 2 (Xu & Mannor (2012))
Étant donné un échantillon d'apprentissage LS de d_l exemples i.i.d. selon un domaine P, soit $M \in \mathbb{N}$, $\epsilon(.) : (X \times Y)^{d_l} \mapsto \mathbb{R}$, un algorithme $\mathcal{A}$ est $(\mathbf{M}, \boldsymbol{\epsilon}(\mathbf{LS}), \hat{\mathbf{d}}_\mathbf{l}(\mathbf{LS}))$ **pseudo-robuste sur P**, *pour $\hat{d}_l(.) : (X \times Y)^{d_l} \mapsto \mathbb{R}$, si $X \times Y$ est partitionnable en M ensembles disjoints, notés par $\{C_i\}_{i=1}^M$, tels qu'il existe un sous-ensemble $\hat{LS} \subseteq LS$, avec $|\hat{LS}| = \hat{d}_l(LS)$ pour tout exemple* $\mathbf{s}$ *appartenant à $\hat{LS}$, $\mathbf{s}, \mathbf{u} \in C_i \Rightarrow \left|L(\mathcal{A}_{LS}, \mathbf{s}) - L(\mathcal{A}_{LS}, \mathbf{u})\right| \leq \epsilon(LS)$. Avec $\mathcal{A}_{LS}$ le modèle appris depuis LS par $\mathcal{A}$ et $L(\cdot,\cdot)$ la fonction perte de $\mathcal{A}$.*
Si $\hat{d}_l(LS) = d_l$, alors on dit que $\mathcal{A}$ est $(\mathbf{M}, \boldsymbol{\epsilon}(\mathbf{LS}))$ **robuste sur P**.

Étant donné l'ensemble d'apprentissage LS, la (pseudo-)robustesse d'un algorithme, mesurée par les valeurs de M, $\epsilon(LS)$ et $\hat{d}_l(LS)$, dépend donc de LS. Un algorithme est pseudo-robuste, si seul un sous-ensemble de LS vérifie la condition. La borne en généralisation suivante peut alors être prouvée.

Théorème 1 (Xu & Mannor (2012))
Si l'échantillon d'apprentissage $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ est i.i.d. selon le domaine P et si un algorithme $\mathcal{A}$ est $(M, \epsilon(LS), \hat{d}_l(LS))$ pseudo-robuste sur P, alors pour tout $\delta > 0$, avec une probabilité d'au moins $1 - \delta$,

$$err_P(\mathcal{A}_{LS}) \leq \hat{err}_P(\mathcal{A}_{LS}) + \frac{\hat{d}_l(LS)}{d_l}\epsilon(LS) + L^{up}\left(\frac{d_l - \hat{d}_l(LS)}{d_l} + \sqrt{\frac{2M\ln 2 + 2\ln\frac{1}{\delta}}{d_l}}\right),$$

où $err_P(\mathcal{A}_{LS})$ *et* $e\hat{r}r_P(\mathcal{A}_{LS})$ *sont respectivement les erreurs réelle et empirique sur* P *du modèle* $\mathcal{A}_{LS}$ *appris depuis* LS, $L(\cdot,\cdot)$ *étant bornée par* L^{up}.

Cette borne n'est pas vérifiée dans le contexte de l'AD. Xu & Mannor (2012) ont néanmoins discuté l'existence d'une telle borne basée sur une divergence spécifique entre les domaines. Nous proposons ici une borne sur l'erreur cible pour (*SSDASF*$_{opt}$) en considérant la $\mathcal{H}\Delta\mathcal{H}$-distance de la borne d'AD (1) plus adaptée à notre cadre. Nous prouvons que (*SSDASF*$_{opt}$) est pseudo-robuste sur les domaines, puis nous déduisons la borne en généralisation suivante.

Théorème 2
Soit (X,ρ) *un espace métrique compact,* K *une bonne fonction de similarité continue sur son premier argument et* $\mathcal{H}$ *la classe des classifieurs-SF. Soit les hyperparamètres* $\beta > 0$, $\lambda > 0$, $\kappa \in [0,1]$, *l'ensemble de landmarks* R *et un ensemble de paires* $\mathcal{C}_{ST}$ *tel que* $B_R > 0$. *Si* LS *contient* $(1-\theta)d_l$ *exemples i.i.d. selon* P_S *le domaine source et* θd_l *exemples i.i.d. selon* P_T *le domaine cible, alors* (*SSDASF*$_{opt}$) *est* $\left(\mathbf{2M}_\eta, \frac{\mathbf{N}_\eta}{(\mathbf{1}-\kappa)\beta\mathbf{B_R}+\lambda}, (\mathbf{1}-\theta)\mathbf{d_l}\right)$ **pseudo-robuste sur** $\mathbf{P_S}$ *et* $\left(\mathbf{2M}_\eta, \frac{\mathbf{N}_\eta}{(\mathbf{1}-\kappa)\beta\mathbf{B_R}+\lambda}, \theta\mathbf{d_l}\right)$ **pseudo-robuste sur** $\mathbf{P_T}$ *où* $\eta > 0$, M_η *est le nombre d'*η*-couvertures de* X, $N_\eta = \max\left\{\max_{\substack{\mathbf{x}_a,\mathbf{x}_b\sim D_S\\ \rho(\mathbf{x}_a,\mathbf{x}_b)\leq\eta}} \|{}^t\phi^R(\mathbf{x}_a) - {}^t\phi^R(\mathbf{x}_b)\|_\infty, \max_{\substack{\mathbf{x}_a,\mathbf{x}_b\sim D_T\\ \rho(\mathbf{x}_a,\mathbf{x}_b)\leq\eta}} \|{}^t\phi^R(\mathbf{x}_a) - {}^t\phi^R(\mathbf{x}_b)\|_\infty\right\}$. *Pour tout* $h \in \mathcal{H}$ *appris avec* (*SSDASF*$_{opt}$), *si* $h^* = \operatorname{argmin}_{h'\in\mathcal{H}}\left\{err_T(h') / e\hat{r}r_\kappa(h) \leq e\hat{r}r_\kappa(h')\right\}$, *alors pour tout* $\delta > 0$, *avec une probabilité d'au moins* $1-\delta$, *on a,*

$$err_T(h) \leq err_T(h^*) + \sqrt{\frac{\kappa^2}{\theta} + \frac{(1-\kappa)^2}{1-\theta}}\sqrt{\frac{\log\frac{4}{\delta}}{2d_l}} + \frac{(2\theta-1)N_\eta\kappa + (1-\theta)N_\eta}{(1-\kappa)\beta B_R + \lambda}$$
$$+\theta + \kappa(1-2\theta) + \sqrt{\frac{4M_\eta \ln 2 + 2\ln\frac{4}{\delta}}{d_l}} + 2(1-\kappa)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu\right).$$

Démonstration. Voir Morvant *et al.* (2012). ■

Dans cette borne, $d_{\mathcal{H}\Delta\mathcal{H}}$ et ν mesurent la divergence entre les domaines ainsi que la capacité d'adaptation de $\mathcal{H}$. Étant donné le classifieur appris h, $err_T(h^*)$ correspond à l'erreur cible minimale que l'on peut espérer obtenir avec un classifieur d'erreur pondérée empirique au mieux aussi faible que celle de h. La constante $\frac{(2\theta-1)N_\eta\kappa+(1-\theta)N_\eta}{(1-\kappa)\beta B_R+\lambda}$ dépend clairement des termes de régularisation et de N_η. Ce dernier peut être aussi petit que désiré [1], il impliquera alors

1. En choisissant η faible et par continuité de la fonction K sur son premier argument.

une croissance de M_η (converge en $O(1/\sqrt{d_l})$). Une valeur élevée de β, λ ou B_R (*i.e.* les domaines sont éloignés), induit la necessité de plus d'exemples. Si $\kappa = 0$, *i.e.* les données cibles sont ignorées, la borne devient alors une borne dans le cas non-sup. de ($DASF_{opt}$). Si $\kappa = 1$, *i.e.* les données sources sont ignorées, la borne est alors une borne de robustesse dans un cas d'apprentissage supervisé sur les données cibles. En illustration des propriétés de cette borne, nous en proposons une analyse théorique simplifiée par les hypothèses suivantes. Si $\kappa \in [0,a]$ où $a \in]0,1[$ et tend vers 1, alors on a : $\frac{(2\theta-1)\kappa+(1-\theta)N_\eta}{(1-\kappa)\beta B_R+\lambda} < \frac{(2\theta-1)\kappa+(1-\theta)N_\eta}{(1-a)\beta B_R+\lambda}$. Puis, nous notons, $\mathbf{A} = \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)+\nu$, $\mathbf{B} = (1-a)\beta B_R+\lambda$, $\mathbf{C} = (2\theta-1)\left(\frac{N_\eta}{\mathbf{B}}-1\right)-2\mathbf{A}$ et $\mathbf{D}$ les constantes restantes. La borne en généralisation du Th.2 peut alors se réécrire,

$$f(\kappa) = \sqrt{\frac{\kappa^2}{\theta} + \frac{(1-\kappa)^2}{1-\theta}}\sqrt{\frac{\log\frac{4}{\delta}}{2d_l}} + \mathbf{C}\kappa + \theta + \mathbf{D}.$$

Si $d_l^T = \theta d_l$ et $d_l^S = (1-\theta)d_l$, la forme de la racine maximale de la dérivée f' est,

$$r = \theta\left(1 + \frac{1-\theta}{\sqrt{\frac{2\log\frac{2}{\delta}}{d_l\mathbf{C}^2} - \theta(1-\theta)}}\right) = \frac{d_l^T}{d_l^S + d_l^T}\left(1 + \frac{d_l^S}{\sqrt{\frac{2\log\frac{2}{\delta}(d_l^S+d_l^T)}{\mathbf{C}^2} - d_l^S d_l^T}}\right).$$

Si η est tel que $N_\eta = \mathbf{B}$ et d_l est suffisamment élevé, alors $\mathbf{C}^2 = 4\mathbf{A}^2$. Donc, r existe et est valide si $d_l^T < \frac{\log(4/\delta)}{(1-\theta)2\mathbf{A}^2}$ et la valeur optimale de κ est définie par,

$$\kappa = \begin{cases} a & \text{si } d_l^T \geq \frac{\log(4/\delta)}{(1-\theta)2\mathbf{A}^2}, \\ \min\{a, r\} & \text{si } d_l^T < \frac{\log(4/\delta)}{(1-\theta)2\mathbf{A}^2}. \end{cases}$$

Si $\mathbf{A} = 0$, la borne suggère $\kappa = \theta$: si les domaines sont indiscernables, alors κ doit suivre la répartition de LS. Dans le cas non-sup. avec $d_l^T = 0$, alors la borne suggère $\kappa = 0$. Si l'on ne considère que des données cibles avec $d_l^S = 0$, alors $\kappa = 1$, ce qui est cohérent dans notre cadre. Si les domaines sont très différents, *i.e.* $\mathbf{A}$ est maximal, alors κ tend vers 1 : seules les données cibles seront pertinentes car la tâche est trop difficile. Cette tendance est confirmée par B_R : si les distributions sont éloignées dans le ϕ^R-espace, alors B_R sera grand et affecter un poids plus élevé aux données cibles sera plus judicieux. Si $\mathbf{B}$ est faible, alors les paramètres sont petits : le poids sur la minimisation de la distance est faible et/ou des modèles complexes sont autorisés. Un tel cas suggère un κ élevé afin de plutôt se concentrer sur les données cibles. Dans le

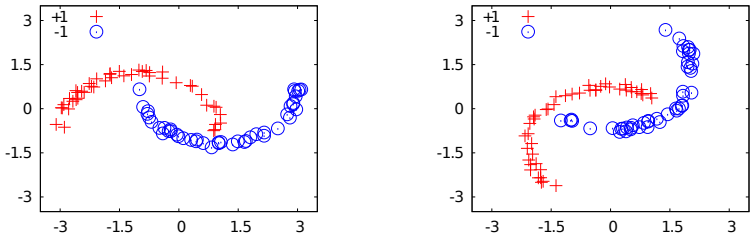


FIGURE 1: **(Jouet)** Droite : domaine source. Gauche : domaine cible ($50°$).

cas contraire, un κ faible pourra être une meilleure solution. En conclusion, cette analyse est proche de celle de Ben-David *et al.* (2010a) mais a l'avantage de considérer la régularisation pour expliquer le comportement de l'approche.

6. Étude empirique du comportement

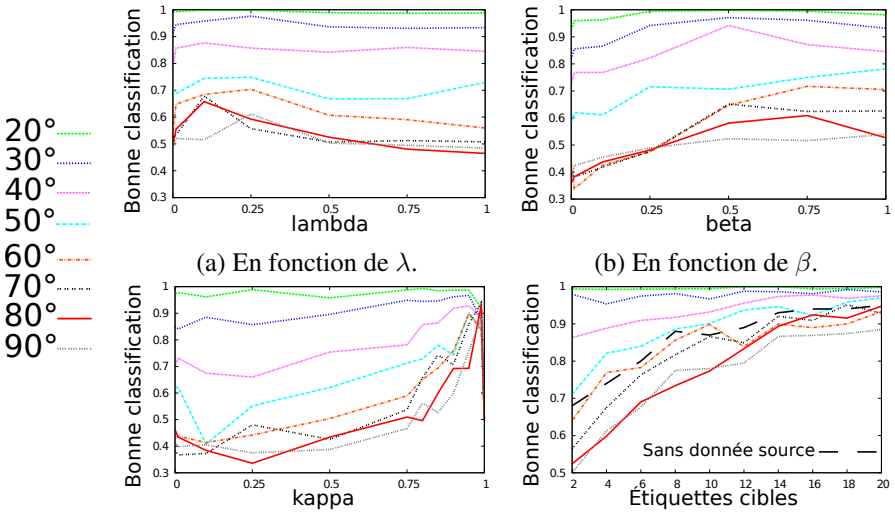
L'objectif est d'effectuer une étude expérimentale de la méthode et de confirmer les propriétés théoriques indiquées dans la Sec.5. En effet, Morvant *et al.* (2011) ont déjà évalué avec succès les performances de DASF.

Protocole expérimental. L'ensemble de paires $\mathcal{C}_{ST}$ et la fonction de similarité [2] sont choisis par validation inverse (*cf.* Morvant *et al.* (2011)). Dans un premier temps, nous évaluons l'impact des paramètres, λ (avec β, κ fixés) et β (avec λ, κ fixés). Les valeurs testées sont : 0, $.01$, $.1$, $.25$, $.5$, $.75$, 1. Afin de s'affranchir de l'influence des étiquettes cibles, nous fixons $\kappa = 0$ (les comportements restent similaires si $\kappa > 0$). Dans un second temps, nous étudions la capacité de SSDASF à apprendre un bon classifieur si une partie de l'échantillon étiqueté est issue du domaine cible. Tout d'abord, à λ, β et $d_l^T = 10$ fixés, nous étudions l'impact de κ avec les valeurs : 0, $.01$, $.1$, $.25$, $.5$, $.75$, $.80$, $.85$, $.90$, $.95$, $.99$, 1. Enfin, nous testons différentes quantités d_l^T de données cibles étiquetées : 2, 4, 8, 10, 12, 14, 16, 18, 20. λ, β et κ sont alors choisis par validation inverse. Dans ce cas, nous nous comparons à un classifieur-SF appris à partir des données cibles ("Sans donnée source" sur les figures).

6.1. Problème jouet : Les lunes jumelles

Le domaine source correspond à un problème de classification binaire classique dît de lunes jumelles (une classe par lune, *cf.* Fig.1). Nous considérons 8

2. Le choix se fait entre un noyau Gaussien et une normalisation non-SDP de ce noyau.



(a) En fonction de λ. (b) En fonction de β.

(c) En fonction de κ. (d) En fonction de d_l^T.

FIGURE 2: **(Jouet)** Taux moyens de bonne classification.

TABLE 1: **(Jouet)** Quantité moyenne de landmarks illustrant la parcimonie. En gras, les valeurs associées au modèle le plus performant.

(a)En fonction de λ.

λ	0	.01	.1	.25	.5	.75	1
20°	16	15	**10**	15	18	13	15
30°	12	10	**9**	11	17	15	21
40°	13	9	**8**	12	14	13	16
50°	8	6	**6**	5	10	10	9
60°	12	5	3	**4**	6	8	12
70°	6	5	**3**	5	6	9	12
80°	7	8	**2**	6	9	8	8
90°	8	11	6	**3**	4	8	7

(b)En fonction de β.

β	0	.01	.1	.25	.5	.75	1
20°	24	22	16	13	**11**	12	14
30°	24	18	17	11	**10**	16	11
40°	24	20	17	13	**8**	11	8
50°	24	20	19	12	11	11	**6**
60°	22	4	3	3	3	**3**	3
70°	20	6	4	3	**3**	4	5
80°	20	6	5	5	3	**3**	4
90°	20	11	9	7	7	6	**5**

(c) En fonction de κ.

κ	0	.01	.1	.25	.5	.75	.8	.85	.9	.95	.99
20°	24	18	22	21	19	19	**22**	24	24	21	7
30°	22	19	18	19	18	17	20	18	21	**23**	7
40°	16	11	13	14	19	16	10	24	18	**20**	7
50°	18	17	10	17	11	20	16	18	23	**14**	7
60°	11	15	11	10	12	9	12	7	14	**11**	12
70°	8	7	9	17	8	14	14	14	12	12	**9**
80°	17	15	4	6	13	8	9	13	9	16	**11**
90°	8	6	11	11	10	14	16	11	10	14	**7**

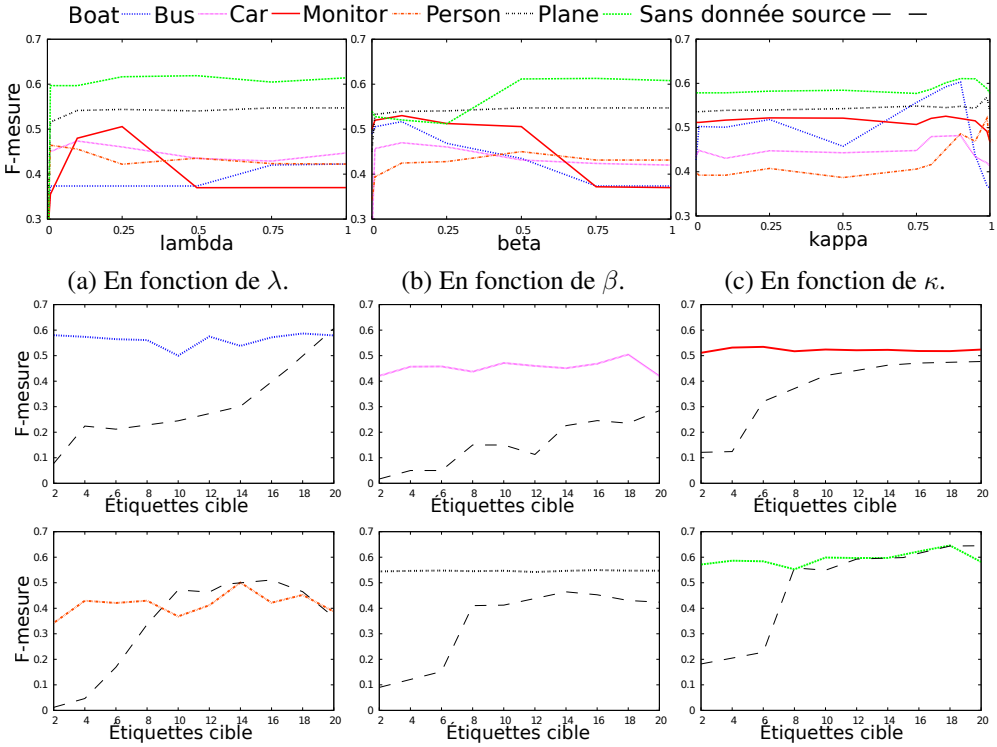
domaines cibles différents associés à une rotation anti-horaire (selon 8 angles) du domaine source. Plus l'angle est grand, plus la tâche d'AD est difficile (*i.e.* plus B_R tend a être élevé). Nous générons 10 tirages de 300 instances ($150+$, $150-$) par domaine et un échantillon test de 1500 exemples cibles.

Influence de λ et β. Dans un premier temps, les Figs.2(a)(b) reportent le taux moyen de bonne classification pour chaque angle de rotation en fonction respectivement de λ (selon le meilleur β) et de β (selon le meilleur λ). Nous observons un impact de β plus significatif que celui de λ. En effet, une bonne valeur de β mène à un gain de $.05$ à $.35$, alors que le gain associé à un λ approprié ne dépasse pas $.2$. Les λ et β menant aux modèles les plus performants appartiennent respectivement à $[.1,.25]$ et $[.5,1]$. Notons que les tâches

dures sont plus sensibles aux paramètres (pour les plus simples, le gain avec λ est *quasi* nul). Dans un second temps, les Tabs.1(a)(b) indiquent le nombre moyen de landmarks associé aux modèles précédemment considérés. Ici, c'est la combinaison de λ et β qui infère des modèles plus parcimonieux. De plus, comme le suggère le Lem.1, la parcimonie augmente avec la valeur de β associée au modèle le plus performant et avec la difficulté de la tâche (*cf.* Tab.1(b)).
Influence de données d'apprentissage cibles. Tout d'abord, le taux moyen de bonne classification en fonction de la quantité de données d'apprentissage cibles d_l^T est indiqué sur la Fig.2(d). On y observe le comportement attendu de l'augmentation de la performance avec d_l^T. Plus la tâche est dure, plus cette augmentation est significative. Cependant, pour les tâches les plus difficiles ($\geq 70°$) nous sommes incapables d'inférer un classifieur plus efficace que le classifieur-SF appris uniquement depuis les étiquettes cibles. Ce résultat reste cohérent avec la borne en généralisation du Th.2, puisque les tâches dures requièrent plus d'étiquettes cibles si les domaines sont éloignés, et parfois se focaliser uniquement sur ces données cibles sera plus judicieux. Ensuite, la Fig.2(c) reporte le comportement en fonction de κ (selon les meilleurs λ et β et avec $d_l^T = 10$). Comme attendu, un κ élevé, entre $.9$ et $.99$, est préféré : le gain est entre $.1$ et $.5$. Encore une fois, l'impact est plus significatif pour les tâches complexes. Pour les plus simples, le Tab.1(c) montre que l'influence de κ sur la parcimonie est directe mais est amoindrie lorsque B_R tend à être élevé. Enfin, comme espéré par le Lem.1, considérer des étiquettes cibles implique des modèles moins parcimonieux.

6.2. Problème réel : Classification d'images

Nous réalisons une AD du corpus PascalVOC'07 (Everingham *et al.*, 2007) vers le corpus TrecVid'07 (Smeaton *et al.*, 2009). Le but est d'identifier des objets (*concepts*) visuels classiques dans des images. PascalVOC est constitué de 5000 images d'apprentissage et de 5000 de test, TrecVid d'images extraites de vidéos. Les images sont représentées par un descripteur visuel proposé par Ayache *et al.* (2007) et défini par les scores de prédictions sur 15 concepts "intermédiaires" (ANIMAL, BUILDING, CAR, CARTOON, EXPLOSION-FIRE, FLAG-US, GREENERY, MAPS, ROAD, SEA, SKIN_FACE, SKY, SNOW, SPORTS, STUDIO_SETTING) détectés par des classifieurs-SVM appris à partir de moments couleurs et d'orientations de contours sur 260 blocs de 32×32 pixels (dimension de 3900). Nous considérons les 6 concepts communs aux corpus : BOAT, BUS, CAR, MONITOR, PERSON, PLANE. Nous générons un échantillon source issu de PascalVOC



(a) En fonction de λ. (b) En fonction de β. (c) En fonction de κ.

(d) En fonction de d_l^T.

FIGURE 3: **(Images)** Résultats en terme de F-mesure moyenne.

et constitué de tous les exemples d'apprentissage positifs et d'exemples négatifs tirés tel que le taux de positif soit de 33%. L'échantillon cible, avec la même répartition, est issu de TrecVid. Les performances sont alors évaluées avec la F-mesure classique : $\frac{2.Précision.Rappel}{Précision+Rappel}$.

Influence de λ et β. Sur les Figs.3(a)(b) nous observons respectivement l'impact de λ (selon le meilleur β) et de β (selon le meilleur λ). La Fig.3(a) montre une influence relative de λ, excepté pour CAR où la meilleur F-mesure est proche de $.25$ avec un gain d'au moins $.15$. Pour les autres concepts, le gain est inférieur à $.1$ et le λ le plus approprié, strictement positif, dépend du problème. À l'instar du problème jouet, la Fig.3(b) montre clairement un impact de β majoritairement plus important : pour BOAT, CAR et PLANE le gain est de $.1$ à $.15$. Ainsi, en se focalisant plus finement sur β, la recherche des paramètres peut être allégée. La valeur de β menant au meilleur classifieur, strictement positive, appartient à $[.01,.25]$, sauf PLANE préfèrant un β supérieur à $.5$.

Influence de données d'apprentissage cibles. Tout d'abord, la Fig.3(d) reporte les résultats pour chaque concept. Ils améliorent ceux sans étiquette cible (avec $\kappa = 0$). De plus, lorsque $d_l^T < 8$, les modèles trouvés sont toujours plus performants que le classifieur-SF appris uniquement à partir des étiquettes cibles. Plus de 20 exemples sont parfois nécessaires pour atteindre les performances de SSDASF, montrant l'utilité des étiquettes cibles. Ensuite, nous avons testé κ (selon les meilleurs λ et β et avec $d_l^T = 10$). Les résultats sont reportés sur la Fig.3(c). Les valeurs les plus pertinentes pour κ sont clairement entre $.9$ et $.99$. Pour ce jeu de données plus difficile, les données cibles apportent donc une information importante. Finalement, le descripteur utilisé ne semble pas être assez expressif, expliquant la difficulté à obtenir de meilleurs résultats. En effet, en multimédia, les données sont souvent représentées en fonction de plusieurs modalités pour permettre une plus grande expressivité, une perspective serait alors d'adapter SSDASF à la multi-modalité.

7. Conclusion

Un bon algorithme d'Adaptation de Domaine doit être capable de tirer parti de toute l'information cible disponible et *a fortiori* d'étiquettes cibles. Avec cette idée, nous avons proposé une généralisation de DASF (Morvant *et al.*, 2011, 2012) au cadre de l'AD semi-supervisée, permettant de considérer l'ensemble des données étiquetées disponibles. Comme DASF, notre méthode - SSDASF - est capable d'inférer des modèles parcimonieux même lorsque la tâche d'AD est difficile. Nous avons, de plus, prouvé que notre programme linéaire est pseudo-robuste (Xu & Mannor, 2012), ce qui nous a permis d'étudier la capacité en généralisation de notre méthode en la liant aux termes de régularisation. Notre étude théorique a été confirmée empiriquement lors des expériences réalisées.

Remerciements : Travail financé par le projet ANR VideoSense ANR-09-CORD-026.

Références

AYACHE S., QUÉNOT G. & GENSEL J. (2007). Image and video indexing using networks of operators. *Journal on Image and Video Processing*.

BALCAN M., BLUM A. & SREBRO N. (2008). Improved guarantees for learning via similarity functions. In *Proceedings of COLT*.

BEN-DAVID S., BLITZER J., CRAMMER K., KULESZA A., PEREIRA F. & VAUGHAN J. (2010a). A theory of learning from different domains. *Machine Learning Journal*, **79**.

BEN-DAVID S., LU T., LUU T. & PAL D. (2010b). Impossibility theorems for domain adaptation. *JMLR W&CP*.

BERGAMO A. & TORRESANI L. (2010). Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Proceedings of NIPS*.

BLITZER J., FOSTER D. & KAKADE S. (2011). Domain adaptation with coupled subspaces. In *Proceedings of AISTATS*.

DAUMÉ III H. (2007). Frustratingly easy domain adaptation. In *Proceedings of ACL*.

DAUMÉ III H., KUMAR A. & SAHA A. (2010). Co-regularization based semi-supervised domain adaptation. In *Proceedings of NIPS*.

EVERINGHAM M., VAN GOOL L., WILLIAMS C. K. I., WINN J. & ZISSERMAN A. (2007). The PASCAL visual object classes challenge 2007 (voc 2007) results. `www.pascal-network.org/challenges/VOC/voc2007/`.

HUANG J., SMOLA A., GRETTON A., BORGWARDT K. & SCHÖLKOPF B. (2006). Correcting sample selection bias by unlabeled data. In *Proceedings of NIPS*.

JIANG J. (2008). *A Literature Survey on Domain Adaptation of Statistical Classifiers*. Rapport interne, Comp. Sc. Dep. at Univ. of Illinois. `sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf`.

JIANG J. & ZHAI C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of ACL*.

MANSOUR Y., MOHRI M. & ROSTAMIZADEH A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*.

MORVANT E., HABRARD A. & AYACHE S. (2011). Sparse domain adaptation in projection spaces based on good similarity functions. In *Proceedings of ICDM*.

MORVANT E., HABRARD A. & AYACHE S. (2012). Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems (KAIS)*. To appear.

PAN S. & YANG Q. (2010). A survey on transfer learning. *IEEE TKDE*.

QUIONERO-CANDELA J., SUGIYAMA M., SCHWAIGHOFER A. & LAWRENCE N. (2009). *Dataset Shift in Machine Learning*. MIT Press.

SMEATON A., OVER P. & KRAAIJ W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*. Springer Verlag.

SUGIYAMA M., NAKAJIMA S., KASHIMA H., VON BÜNAU P. & KAWANABE M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of NIPS*.

XU H. & MANNOR S. (2012). Robustness and generalization. *Machine Learning Journal*, **86**(3), 391–423.