Un Cadre Formel de Boosting pour l'Adaptation de Domaine

Amaury Habrard, Jean-Philippe Peyrache, Marc Sebban

Un Cadre Formel de Boosting pour l'Adaptation de Domaine

Amaury Habrard, Jean-Philippe Peyrache, Marc Sebban

Université de Saint-Etienne, Université de Lyon
Laboratoire Hubert Curien, UMR CNRS 5516,
18 rue du Pr Benoit Lauras, 42000 Saint-Etienne
firstname.lastname@univ-st-etienne.fr

Résumé : L'hypothèse PAC classique, selon laquelle les données d'apprentissage et de test sont issues d'une même distribution, n'est pas satisfaite dans bon nombre de problèmes réels en apprentissage. Pour traiter des cas où les données sont issues de distributions source et cible différentes, un cadre récent - l'adaptation de domaine (AD) - permet de concevoir des algorithmes théoriquement fondés. Nous présentons DABOOST basé à la fois sur les théories du Boosting et de l'AD. L'originalité par rapport à l'état de l'art est que notre méthode nécessite uniquement des étiquettes source. Aucune information sur les étiquettes cible n'étant disponible, nous minimisons à la fois l'erreur en classification sur la source, comme ADABOOST, et la proportion de violations de marge sur la cible. Nous montrons théoriquement la convergence de DABOOST avec une borne sur l'erreur en généralisation et donnons des exemples pratiques de son efficacité. **Mots-clés** : Boosting, adaptation de domaine, données cibles non étiquetées

1. Introduction

La plupart des algorithmes d'apprentissage se basent sur l'hypothèse selon laquelle les données d'entraînement sont issues de la même distribution que les données test. Cependant, ce postulat n'est que très rarement vérifié dans beaucoup d'applications réelles, remettant en question les théories classiques de l'apprentissage comme le modèle PAC (Valiant, 1984). Pour contourner de tels problèmes, un nouveau cadre a récemment été introduit, menant à l'émergence de la théorie de *l'adaptation de domaine* (AD) (Ben-David *et al.*, 2010; Mansour *et al.*, 2009). Une situation typique d'AD peut être décrite de la manière suivante : l'algorithme d'apprentissage reçoit des données étiquetées du domaine *source* (ou éventuellement de plusieurs sources (Mansour *et al.*,

2008)), et des points de la distribution dite *cible*, peu ou pas étiquetés. On retrouve l'AD dans de nombreuses applications (Martínez, 2002; Roark & Bacchiani, 2003; Blitzer *et al.*, 2007; Chelba & Acero, 2006).

Ces dernières années, de nouveaux résultats fondamentaux en AD ont ouvert la porte à la conception d'algorithmes d'AD théoriquement fondés. Dans cet article, nous nous concentrons sur le scénario selon lequel l'ensemble d'apprentissage est composé de données étiquetées de la source et d'exemples *non étiquetés* de la cible. Pour traiter cette situation plus complexe, nous présentons un algorithme d'AD innovant, qui tire son origine à la fois de la théorie du boosting (Freund & Schapire, 1996), ainsi que de celle de l'AD. Nous affirmons que le boosting convient particulièrement à des problèmes d'AD pour les raisons suivantes : (i) le boosting (via son fameux algorithme ADABOOST) est par nature une procédure *adaptative*, dans la mesure où les exemples d'apprentissage sont repondérés afin de créer de la diversité dans les hypothèses apprises (ii) on sait que le boosting maximise la marge (Schapire *et al.*, 1997). Ce dernier point apparait donc très utile dans le cas où l'ensemble d'apprentissage contient également des données non étiquetées, pour lesquelles le taux d'erreur empirique ne peut pas être optimisé.

Le boosting a déjà été exploité dans des méthodes d'AD, mais uniquement dans des situations dans lesquelles l'algorithme d'apprentissage reçoit des données cible étiquetées. TRADABOOST (Dai *et al.*, 2007), par exemple, réutilise le schéma de pondération d'ADABOOST sur les exemples cible. Des approches existent sans utilisation d'étiquettes cible, mais se situent plutôt dans un cadre semi-supervisé (*i.e.* les domaines source et cible sont les mêmes).

Dans cet article, notre objectif est de faire le lien entre l'AD et le boosting, dans le cas où **les données cible dont on dispose sont non étiquetées** (en plus de données source étiquetées). Nous présentons DABOOST, qui optimise à la fois *le taux d'erreur en classification* sur les exemples source étiquetés, ainsi que la *proportion de violations de marge* (par rapport à une marge γ) sur les points cible non étiquetés. La Figure 1 montre le principe sous-jacent de notre approche, basé sur le fait que plus la distance d'un exemple non étiqueté au classifieur est grande, plus la probabilité qu'on lui attribue la bonne classe est importante. Cependant, il est intéressant de noter que ce principe n'est valide que dans le cas où le classifieur obtient des marges pertinentes, *i.e.* uniquement s'il est efficace sur les données source. Par conséquent, l'objectif de DABOOST est de réduire la proportion de points *cible* situés à une distance plus petite que γ du classifieur, tout en minimisant le taux d'erreur en classification sur les exemples *source*. Il faut également mentionner que

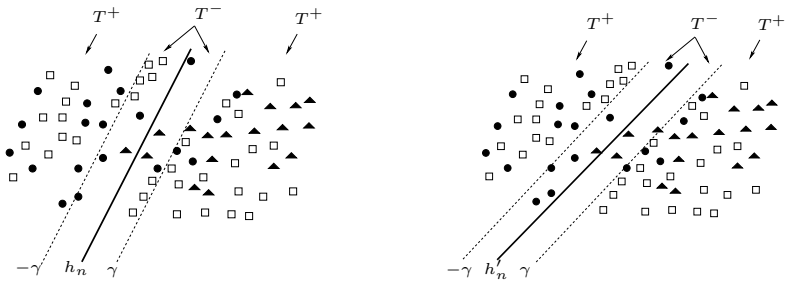


FIGURE 1: (a) : Les exemples en noir représentent les données source étiquetées (cercle ou triangle). Les carrés représentent les points cibles non étiquetés. Notre but est de réduire le nombre de points cible dans T^-, tout en minimisant le taux d'erreur en classification sur la source. Les exemples cible dans T^+ sont ceux pour lesquels on a une confiance importante, car leur marge est supérieure à γ. (b) : Résultat de l'adaptation : h'_n satisfait la contrainte de marge, tout en conservant une erreur faible sur les exemples source.

l'idée de satisfaire une marge donnée sur les exemples cible a aussi été exploitée dans DASVM (Bruzzone & Marconcini, 2010), auquel DABOOST va être comparé dans ce papier. Cependant DASVM, qui est une approche basée sur la théorie des SVM plutôt que sur le boosting, ne fournit aucune garantie en généralisation sur l'erreur cible. Dans cet article, après avoir défini la notion d'apprenant faible pour l'AD, nous prouvons que (i) à la fois l'erreur sur la source et le risque d'avoir de faibles marges sur la cible, tendent vers 0 avec les itérations de DABOOST et (ii) ce sont deux conditions nécessaires afin de diminuer l'erreur en généralisation sur la distribution cible. Nous dérivons également une nouvelle borne en généralisation, qui prend en compte un concept innovant de divergence entre deux distributions. Nous montrons expérimentalement que dans notre cadre, minimiser cette borne plutôt que celle de Ben-David *et al.* (2010), permet de trouver un meilleur classifieur final.

La suite du papier est organisée comme suit : les notations et définitions sont introduites dans la Section 2 ; DABOOST est présenté dans la Section 3 ; les analyses théoriques de notre algorithme se situent dans la Section 4 ; enfin, nous menons une séries d'études expérimentales et montrons son efficacité en pratique dans la Section 5.

2. Définitions et Notations

Définition 1
Soit S un échantillon de données étiquetées (x, y) tirées depuis une distribution source $\mathcal{S}$ sur $X \times \{-1, +1\}$, où X est l'espace de représentation. Soit T un échantillon d'exemples non étiquetés $(x, .)$ tirés depuis une distribution cible $\mathcal{T}$ sur X. Soit $\mathcal{H}$ une classe d'hypothèses et $h_n \in \mathcal{H} : X \to [-1, +1]$ une hypothèse apprise sur $S \cup T$ et la distribution empirique correspondante D_n. Nous définissons les ensembles suivants (voir Figure 1), $\forall\ \gamma \leq 1$:

$$T^- = \{x_i \in T\ t.q.\ |Bh_n(x_i)| \leq \gamma\},\ \textit{and}\ T^+ = \{x_i \in T\ t.q.\ |h_n(x_i)| > \gamma\}.$$

T^- est l'ensemble des points cible situés à l'intérieur d'une bande de marge de taille γ, de chaque côté de h_n. T^+ regroupe les points cible qui ont une marge en valeur absolue plus grande que γ. De ces ensembles, nous déduisons W_{T^+} (resp. W_{T^-}) comme étant la somme des poids des exemples cible appartenant à T^+ (resp. T^-). De plus, W_{S^+} (resp. W_{S^-}) est la somme des poids des points source correctement (resp. incorrectement) étiquetés par l'hypothèse h_n. On notera que $W_{S^+} + W_{S^-} + W_{T^+} + W_{T^-} = 1$.

Définition 2 (Apprenant faible)
Inspirés de Freund & Schapire (1996), nous définissons une hypothèse h_n apprise à une itération n comme un apprenant faible sur un échantillon d'apprentissage étiqueté S, tiré selon $\mathcal{S}$, si h_n a un taux de succès au moins un peu meilleur que l'aléatoire, c'est-à-dire $\exists\ \tau_n^S \in]0; \frac{1}{2}]$:

$$\hat{\epsilon}_n = \hat{Pr}_{x_i \sim D_n^S}[h_n(x_i) \neq y_i] = \frac{W_{S^-}}{W_S} = \frac{1}{2} - \tau_n^S,$$

où $W_S = W_{S^-} + W_{S^+}$, $D_n^S(x_i) = \frac{D_n(x_i)}{W_S}$ si $x_i \in S$, 0 sinon et où $[.]$ est une fonction indicatrice. Nous généralisons cette Définition à l'AD.

Définition 3 (Apprenant faible pour l'AD)
Un classifieur h_n appris à l'itération n depuis un échantillon source étiqueté S tiré selon $\mathcal{S}$ et un échantillon cible non étiqueté T tiré selon $\mathcal{T}$ est un apprenant faible pour l'AD sur T si $\forall\ \gamma \leq 1$:

1. *h_n est un apprenant faible sur S.*
2. *$\exists\ \tau_n^T \in]0; \frac{1}{2}]$ tel que $\hat{L}_n = \hat{Pr}_{x_i \sim D_n^T}[|h_n(x_i)| \leq \gamma] = \frac{W_{T^-}}{W_T} = \frac{1}{2} - \tau_n^T$,*

où $W_T = W_{T^-} + W_{T^+}$, $D_n^T(x_i) = \frac{D_n(x_i)}{W_T}$ si $x_i \in T$, 0 sinon.

La première condition de la Définition (3) signifie que pour être capable d'adapter de $\mathcal{S}$ à $\mathcal{T}$ en suivant un schéma de boosting, h_n doit apprendre quelque de nouveau à chaque itération sur la source. La seconde condition signifie que la somme pondérée des points cible possédant une marge plus grande que γ (*i.e.* W_{T^+}) est un petit peu plus importante que celle des exemples à l'intérieur de la bande de marge $[-\gamma, +\gamma]$ (*i.e.* W_{T^-}).

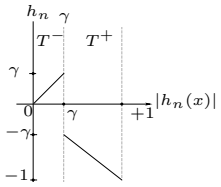
Dans la section suivante, nous présentons DABOOST qui consiste à construire une combinaison d'apprenants faibles pour l'AD.

3. Algorithme DABOOST

Le pseudo-code de DABOOST est présenté dans l'algorithme 1. Démarrant d'une distribution uniforme sur $S \cup T$, il apprend itérativement une nouvelle hypothèse h_n vérifiant les conditions d'apprenant faible pour l'AD sur la distribution D_n de $S \cup T$. Pour réaliser cette tâche, n'importe quelle classe d'hypothèses satisfaisant la Définition (3) peut être utilisée. Cependant, afin de créer de la diversité, nous suggérons de générer aléatoirement un ensemble de stumps sur un attribut sélectionné aléatoirement. De cette manière, on augmente la probabilité de trouver des hypothèses qui satisfont les conditions d'apprenant faible pour l'AD. En effet, le fait d'apprendre uniquement l'hypothèse optimale sur les données source étiquetées (qui satisfait donc la première contrainte de la Définition (3)) augmente le risque de ne pas satisfaire la contrainte sur T. C'est pourquoi DABOOST conserve l'hypothèse qui vérifie le "mieux" les conditions d'apprenant faible pour l'AD[1]. En pratique, nous sélectionnons l'hypothèse qui satisfait au moins $W_{S^+} > W_{S^-}$ (qui est la condition nécessaire pour apprendre le domaine source) et possédant le $W_{T^+} \in]W_{T^-}; b_T]$ le plus grand, ceci afin de s'adapter correctement au domaine cible. On remarquera que (i) b_T borne la valeur de W_{T^+} pour permettre à l'algorithme de générer de la diversité dans la combinaison finale et (ii) il s'agit d'un paramètre à régler en fonction de l'application que l'on traite.

Ensuite, les poids des exemples étiquetés et non étiquetés sont modifiés en fonction de deux règles de mise à jour différentes. Ceux des points source sont mis à jour avec la même stratégie que celle d'ADABOOST. Quant aux points cible, leur poids est modifié selon leur position dans l'espace de représentation. Si un exemple cible x est dans la bande de marge γ, une pseudo-classe

1. Si une telle hypothèse ne peut pas être trouvée, l'adaptation n'est pas possible avec le γ courant, et la procédure de boosting est stoppée. Nous suggérons dans ce cas à l'utilisateur de changer la contrainte de marge, et de relancer à nouveau DABOOST.



Le poids des exemples de T^-, qui ne sont pas à une distance suffisante de h_n, est augmenté de façon monotone en fonction de cette distance. Le poids des points de T^+, dans lesquels on a une grande confiance, est diminué en fonction de leur distance à l'hypothèse.

FIGURE 2: Illustration du comportement des règles de mise à jour.

$y_i = -signe(h_n(x))$ lui est attribuée, simulant une mauvaise classification. Sinon, $y_i = signe(h_n(x))$. De cette manière, notre approche adaptative tend à rendre plus important le rôle joué par les exemples dans la bande de marge, afin de forcer le classifieur suivant à augmenter leur marge. Une illustration de ces règles de mise à jour est représentée dans la Figure 2.

DABOOST peut être vu comme une généralisation d'ADABOOST. En effet, si l'ensemble T est vide, DABOOST n'est rien d'autre qu'ADABOOST. De plus, si T est issu de la distribution source, en d'autres mots si $\mathcal{S} = \mathcal{T}$, DABOOST est une méthode ensembliste semi-supervisée.

Algorithme 1 DABOOST

Entrée : deux ensembles S et T contenant des exemples respectivement étiquetés et non étiquetés ($m = |T| + |S|$), un nombre d'itérations N, une marge $\gamma \leq 1$.
Sortie : un classifieur H_N^S pour la source, et un H_N^T pour la cible
Initialisation : $\forall x_i \in S \cup T, D_1(x_i) = \frac{1}{m}$.
for $n = 1$ **à** N **do**
Apprendre h_n satisfaisant les conditions d'apprenant faible pour l'AD (sinon **sortir de la boucle**).
$\alpha_n = \frac{1}{2} ln \frac{W_{S^+}}{W_{S^-}}$ et $\beta_n = \frac{1}{2\gamma} ln \frac{W_{T^+}}{W_{T^-}}$
Règles de mise à jour :
- $\forall x_i \in S$, $D_{n+1}(x_i) = D_n(x_i).\frac{e^{-\alpha_n signe(h_n(x_i)).y_i}}{Z_n}$.
- $\forall x_i \in T$, $D_{n+1}(x_i) = D_n(x_i).\frac{e^{-\beta_n h_n(x_i).y_i^n}}{Z_n}$,
où $y_i^n = signe(h_n(x_i))$ si $|h_n(x_i)| > \gamma$, $y_i^n = -signe(h_n(x_i))$ sinon,
et Z_n est un coefficient de normalisation.
end for
$f_N^S(x) = \sum_{n=1}^N \alpha_n signe(h_n(x))$ et $f_N^T(x) = \sum_{n=1}^N \beta_n signe(h_n(x))$.
Classifieurs source et cible finaux :
$H_N^S(x) = signe(f_N^S(x))$ et $H_N^T(x) = signe(f_N^T(x))$.

4. Analyse théorique

Dans cette section, nous présentons une analyse théorique de DABOOST. La qualité d'une hypothèse h_n se mesure par sa capacité à classifier correctement non seulement les exemples étiquetés de S mais aussi les exemples non étiquetés de T avec une marge importante. En supposant que les contraintes d'apprenant faible pour l'AD de la Définition(3) sont satisfaites, incluant la condition d'apprenant faible sur S, on peut noter que tous les résultats classiques d'ADABOOST sont toujours valables sur $\mathcal{S}$. Ce qui signifie que l'erreur empirique $\hat{\epsilon}_{H_N^S}$ sur S décroît avec N (Schapire *et al.*, 1997). Par la suite, nous montrons que la perte $\hat{L}_{H_N^T}$, qui représente la proportion d'exemples cible possédant une marge plus petite que γ après N itérations, diminue elle aussi avec N. Puis, nous montrons qu'une telle décroissance, associée à la réduction d'une mesure de divergence entre les distributions $\mathcal{S}$ et $\mathcal{T}$, nous permet de dériver une borne de l'erreur en généralisation sur $\mathcal{T}$. Nous tirons profit de cette borne afin de proposer un critère d'arrêt pour DABOOST.

4.1. Bornes sur les pertes empiriques

Théorème 1
Soit $\hat{L}_{H_N^T}$ la proportion des exemples cible de T possédant une marge plus petite que γ après N itérations de DABOOST*.*

$$\hat{L}_{H_N^T} = \hat{Pr}_{x_i \sim T}[\mathbf{y}_i \mathbf{f_T^N}(x_i) < 0] \leq \frac{1}{|T|} \sum_{x_i \sim T} e^{-\mathbf{y}_i \mathbf{f_T^N}(x_i)} \leq \frac{m}{|T|} \Pi_n Z_n, \qquad (1)$$

où $\mathbf{y}_i = (y_i^1, \ldots, y_i^n, \ldots, y_i^N)$ est le vecteur des pseudo-classes attribuées à l'exemple cible x_i et $\mathbf{f_T^N}(x_i) = (\beta_1 h_1(x_i), \ldots, \beta_n h_n(x_i), \ldots, \beta_N h_N(x_i))$ est le vecteur des résultats pondérés de DABOOST *pour x_i.*

Preuve Soit $\hat{L}_{H_N^T} = \hat{Pr}_{x_i \sim T}[\mathbf{y}_i \mathbf{f_T^N}(x_i) < 0]$, nous avons :

$$\hat{L}_{H_N^T} = \frac{1}{|T|} \sum_{x_i \sim T} [-\mathbf{y}_i \mathbf{f_T^N}(x_i) \geq 0] \leq \frac{1}{|T|} \sum_{x_i \sim T} e^{-\mathbf{y}_i \mathbf{f_T^N}(x_i)}. \qquad (2)$$

Donc, la première inégalité du Théorème 1 est vérifiée. De plus, $\forall x_i \in T$,

$$D_{N+1}(x_i) = \frac{D_N(x_i) e^{-\beta_N h_N(x_i) y_i^N}}{Z_n} = \frac{D_1(x_i) \Pi_n e^{-\beta_n h_n(x_i) y_i^n}}{\Pi_n Z_n}.$$

Considérant la distribution uniforme à l'itération $n = 1$, on obtient

$$\sum_{x_i \sim T} D_{N+1}(x_i)\Pi_n Z_n = \frac{1}{m}\sum_{x_i \sim T} e^{-\sum_n \beta_n h_n(x_i) y_i^n} = \frac{1}{m}\sum_{x_i \sim T} e^{-\mathbf{y}_i \mathbf{f_T^N}(x_i)}. \quad (3)$$

Nous déduisons des Equations (2) et (3) le dernier résultat du théorème :
$\hat{L}_{H_N^T} \leq \frac{m}{|T|} \times \frac{1}{m}\sum_{x_i \sim T} e^{-\mathbf{y}_i \mathbf{f_T^N}(x_i)} = \frac{m}{|T|}\sum_{x_i \sim T} D_{N+1}(x_i)\Pi_n Z_n \leq \frac{m}{|T|}\Pi_n Z_n$. □

Théorème 2
Soit $\hat{\epsilon}_{H_N^S}$ l'erreur empirique obtenue par le classifieur H_N^S renvoyé par DA-BOOST *après N itérations sur l'échantillon S des exemples source étiquetés.*

$$\hat{\epsilon}_{H_N^S} = \hat{Pr}_{x_i \sim S}[y_i f_N^S(x_i) < 0] \leq \frac{1}{|S|}\sum_{x_i \sim S} e^{-y_i f_N^S(x_i)} \leq \frac{m}{|S|}\Pi_n Z_n.$$

Preuve La preuve est la même que celle du Théorème 1. A l'exception de la constante $\frac{m}{|S|}$, ce Théorème est le même que celui de Freund & Schapire (1996). □

4.2. Coefficients de confiance optimaux

Les Théorèmes 1 et 2 suggèrent la minimisation de chaque Z_n afin de réduire non seulement l'erreur empirique $\hat{\epsilon}_{H_N^S}$ sur S mais aussi la perte empirique $\hat{L}_{H_N^T}$ sur T. A ces fins, nous réécrivons Z_n comme suit :

$$\begin{aligned} Z_n &= \sum_{x_i \in S} D_n(x_i) e^{-\alpha_n signe(h_n(x_i)) y_i} \quad (4) \\ &+ \sum_{x_i \in T^-} D_n(x_i) e^{-\beta_n h_n(x_i) y_i^n} + \sum_{x_i \in T^+} D_n(x_i) e^{-\beta_n h_n(x_i) y_i^n}. \end{aligned}$$

Comme dans Freund & Schapire (1996), on peut réécrire le premier terme :

$$\sum_{x_i \in S} D_n(x_i).e^{-\alpha_n signe(h_n(x_i)) y_i} = W_{S^+} e^{-\alpha_n} + W_{S^-} e^{\alpha_n}. \quad (5)$$

De plus, les deux derniers termes de l'Equation(4), qui implique les exemples cible, peuvent être bornés comme suit (cf Figure 2 pour une justification) :
- $\forall x_i \in T^-$, $D_n(x_i) e^{-\beta_n h_n(x_i) y_i^n} \leq D_n(x_i) e^{-\beta_n \gamma}$,
- $\forall x_i \in T^+$, $D_n(x_i) e^{-\beta_n h_n(x_i) y_i^n} \leq D_n(x_i) e^{\beta_n \gamma}$.

En insérant l'Equation(5) et les bornes précédentes dans l'Equation(4) :

$$Z_n \leq W_{S^+} e^{-\alpha_n} + W_{S^-} e^{\alpha_n} + W_{T^+} e^{-\beta_n \gamma} + W_{T^+} e^{\beta_n \gamma} = Z'_n. \quad (6)$$

En dérivant la combinaison convexe précédente en fonction successivement de β_n et α_n, puis en annulant les termes, nous obtenons les valeurs optimales de β_n et α_n utilisées dans DABOOST :

$$\frac{\partial Z'_n}{\beta_n} = 0 \quad \Rightarrow \quad W_{T^-} e^{\beta_n \gamma} = W_{T^+} e^{-\beta_n \gamma} \Rightarrow \beta_n = \frac{1}{2\gamma} ln \frac{W_{T^+}}{W_{T^-}}. \quad (7)$$

De la même manière, pour α_n, nous obtenons le même paramètre optimal $\alpha_n = \frac{1}{2} ln \frac{W_{S^+}}{W_{S^-}}$ que celui utilisé dans ADABOOST.

4.3. Convergence des pertes empiriques

Le Théorème suivant montre que la perte empirique $\hat{L}_{H_N^T}$ décroît avec le nombre d'itérations de DABOOST, sous contrainte que les conditions d'apprenant faible pour l'AD soient remplies sur S et T.

Théorème 3
Supposons que DABOOST *génère un ensemble de N hypothèses satisfaisant les conditions d'apprenant faible de la Définition(3). Alors, la borne suivante est valide pour la perte empirique $\hat{L}_{H_N^T}$ du classifieur final H_N^T.*

$$\hat{L}_{H_N^T} \leq \frac{m}{|T|} \Pi_{n=1}^N \left(W_S \sqrt{1 - 4\tau_n^{S^2}} + W_T \sqrt{1 - 4\tau_n^{T^2}} \right). \quad (8)$$

Le terme entre parenthèses étant strictement inférieur à 1, ce Théorème spécifie que la perte empirique $\hat{L}_{H_N^T}$ tend vers 0 avec le nombre d'itérations N.

Preuve Par l'Equation(6), nous savons que $Z_n \leq W_{S^+} e^{-\alpha_n} + W_{S^-} e^{\alpha_n} + W_{T^+} e^{-\beta_n \gamma} + W_{T^+} e^{\beta_n \gamma}$. En remplaçant α_n et β_n par leur expressions, on déduit

$$\begin{aligned} Z_n \leq 2\sqrt{W_{S^-} W_{S^+}} + 2\sqrt{W_{T^-} W_{T^+}} &= 2\sqrt{W_S^2 (\tfrac{1}{4} - \tau_n^{S^2})} + 2\sqrt{W_T^2 (\tfrac{1}{4} - \tau_n^{T^2})} \\ &= W_S \sqrt{1 - 4\tau_n^{S^2}} + W_T \sqrt{1 - 4\tau_n^{T^2}} \\ &= (W_S + W_T) max(\sqrt{1 - 4\tau_n^{S^2}}, \sqrt{1 - 4\tau_n^{T^2}}) \\ &< W_S + W_T = 1. \end{aligned}$$

Insérer la deuxième ligne dans le Théorème 1 permet de terminer la preuve. □

On remarquera que le fait d'insérer la deuxième ligne dans le Théorème 2 nous permet de prouver que l'erreur empirique $\hat{\epsilon}_{H_N^S}$ décroît aussi avec DABOOST. Donc, notre algorithme est capable de faire décroître $\hat{L}_{H_N^T}$ et $\hat{\epsilon}_{H_N^S}$.

4.4. Garanties en généralisation

Dans cette section, nous nous concentrons sur la capacité de DABOOST à réduire l'erreur en généralisation sur le domaine cible $\mathcal{T}$. A cet effet, nous utilisons deux types de bornes sur l'erreur en généralisation : la première (voir le Théorème 4 ci-dessous) provient de la théorie du boosting (Schapire *et al.*, 1997) et a le principal avantage de ne pas dépendre du nombre d'itérations du boosting au niveau du terme de pénalisation ; le second est issu de la théorie de l'AD (Ben-David *et al.*, 2010) et suggère de réduire la divergence entre les deux distributions.

Théorème 4 (Schapire *et al.* (1997))
Soit $\mathcal{H}$ une classe de classifieurs de VC-dimension d_h. $\forall\ \delta > 0$ et $\gamma > 0$, avec une probabilité $1 - \delta$, n'importe quel ensemble de N classifieurs construit depuis un échantillon d'apprentissage S de taille $|S|$ issu d'une distribution $\mathcal{S}$ satisfait l'inégalité suivante sur l'erreur en généralisation $\epsilon_{H_N^S}$:

$$\epsilon_{H_N^S} \leq \hat{Pr}_{x\sim S}(marge(x) \leq \gamma) + \mathcal{O}\left(\sqrt{\frac{d_h}{|S|}\frac{log^2(|S|/d_h)}{\gamma^2} + log(1/\delta)}\right). \qquad (9)$$

Ce fameux Théorème affirme que le fait d'obtenir une marge importante sur l'ensemble d'apprentissage (le premier terme de la partie droite) résulte en une amélioration de la borne sur l'erreur en généralisation. De plus, Schapire et al. (Schapire *et al.*, 1997) ont prouvé qu'avec ADABOOST ce terme décroît exponentiellement vite avec le nombre N de classifieurs. En appliquant le Théorème 4 sur l'erreur cible dans le contexte de DABOOST, on peut déduire :

$$\epsilon_{H_N^T} \quad \leq \quad \hat{\mathcal{L}}_T(H_N) + \mathcal{O}\left(\sqrt{\frac{d_h}{|S|}\frac{log^2(|S|/d_h)}{\gamma^2} + log(1/\delta)}\right), \qquad (10)$$

où $\hat{\mathcal{L}}_T(H_N) = \hat{Pr}_{x_i\sim T}[y_i f_N^T(x_i) \leq \gamma]$.

A la différence de $\hat{Pr}_{x\sim S}(marge(x) \leq \gamma)$ dans le Théorème 4, $\hat{\mathcal{L}}_T(H_N)$ ne peut pas être calculée, car elle utilise l'étiquette cible inconnue y_i. Pour contourner ce problème, nous tirons profit de récents résultats présentés dans Ben-David *et al.* (2010) bornant l'erreur cible par l'erreur source et la $\mathcal{H}$-divergence $d_{\mathcal{H}}(\mathcal{S}, \mathcal{T})$ entre les deux domaines $\mathcal{S}$ et $\mathcal{T}$, où $d_{\mathcal{H}}$ est définie ainsi :

$$d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2sup_{h,h'\in\mathcal{H}}|\epsilon_{\mathcal{S}}(h, h') - \epsilon_{\mathcal{T}}(h, h')| \geq 2|\epsilon_{\mathcal{S}}(h, h') - \epsilon_{\mathcal{T}}(h, h')|,$$

où $\forall\ \mathcal{D}, \epsilon_{\mathcal{D}}(h, h') = E_{x\sim\mathcal{D}}[h(x) \neq h'(x)]$ est la probabilité d'un désaccord **en termes de classification** entre h et h' sur $\mathcal{D}$. Pour simplifier, plus la divergence

est faible, plus l'adaptation de $\mathcal{S}$ à $\mathcal{T}$ est réussie. Comme notre objectif est de borner la perte basée sur la marge $\hat{\mathcal{L}}_T(H_N)$ de l'Equation(10), $d_{\mathcal{H}}(\mathcal{S},\mathcal{T})$ n'est pas pertinente car elle est basée sur des désaccords de classification plutôt que sur des désaccords de marge. Afin de régler ce problème, nous introduisons par la suite la $\mathcal{H}_\gamma$-divergence entre deux distributions $\mathcal{S}$ et $\mathcal{T}$.

Définition 4
Soient deux distributions $\mathcal{S}$ et $\mathcal{T}$ sur le même espace de représentation X, et $\mathcal{H}$ une classe d'hypothèses sur X. $\forall \gamma > 0$, la $\mathcal{H}_\gamma$-divergence entre $\mathcal{S}$ et $\mathcal{T}$ est :

$$d_{\mathcal{H}_\gamma}(\mathcal{S},\mathcal{T}) = 2sup_{h,h' \in \mathcal{H}} |\mathcal{L}_{\mathcal{S}}(h,h') - \mathcal{L}_{\mathcal{T}}(h,h')| \geq 2|\mathcal{L}_{\mathcal{S}}(h,h') - \mathcal{L}_{\mathcal{T}}(h,h')|, \quad (11)$$

où $\forall\, \mathcal{D}$, $\mathcal{L}_{\mathcal{D}}(h,h') = E_{x\sim\mathcal{D}}\left[\frac{yh(x)-\gamma}{yh'(x)-\gamma} \leq 0\right]$ est la probabilité d'un désaccord **en termes de marge** *entre h et h' sur $\mathcal{D}$. Par convention, nous écrivons*

$$\mathcal{L}_{\mathcal{D}}(h,.) = \mathcal{L}_{\mathcal{D}}(h) = E_{x\sim\mathcal{D}}\left[yh(x) - \gamma \leq 0\right], \text{ où } .(x) = \frac{1+\gamma}{y}.$$

$\mathcal{L}_{\mathcal{D}}(h,h')$ peut être estimée par $\hat{\mathcal{L}}_D(h,h') = E_{x\sim D}\left[\frac{yh(x)-\gamma}{yh'(x)-\gamma} \leq 0\right]$, où D est un échantillon fini distribué selon $\mathcal{D}$. Avec l'Equation(10) et la $\mathcal{H}_\gamma$-divergence, on peut dériver le Théorème suivant qui borne l'erreur en généralisation sur $\mathcal{T}$.

Théorème 5
Soit $\mathcal{H}$ une classe de classifieurs de VC-dimension d_h. $\forall\ \delta > 0$ et $\gamma > 0$, avec une probabilité de $1-\delta$, le classifieur final appris après N itérations de DABOOST *sur un échantillon d'apprentissage composé de $|S|$ exemples étiquetés distribués selon $\mathcal{S}$ et $|T|$ points non étiquetés distribués selon $\mathcal{T}$ (avec $m = |S|+|T|$) satisfait l'inégalité suivante sur l'erreur en généralisation $\epsilon_{H_N^{\mathcal{T}}}$:*

$$\epsilon_{H_N^{\mathcal{T}}} \quad \leq \quad \hat{\mathcal{L}}_S(H_N) + \frac{1}{2}\hat{d}_{\mathcal{H}_\gamma}(S,T) + \lambda_* + \mathcal{O}\left(\sqrt{\frac{d_h}{|S|}\frac{log^2(|S|/d_h)}{\gamma^2} + log(1/\delta)}\right),$$

où $\hat{\mathcal{L}}_S(H_N) = \hat{Pr}_{x_i\sim S}[y_i f_N^S(x_i) \leq \gamma]$ et λ_* est la perte de l'hypothèse jointe idéale minimisant la perte de marge combinée sur $\mathcal{S}$ et $\mathcal{T}$.

Preuve On notera $H_* = argmin_{h\in\mathcal{H}}(\epsilon_{h^{\mathcal{S}}} + \epsilon_{h^{\mathcal{T}}})$ l'hypothèse optimale minimisant la somme des erreurs en généralisation sur $\mathcal{S}$ et $\mathcal{T}$. En considérant l'inégalité triangulaire pour les fonctions de perte de classification qui implique que pour n'importe

quelles hypothèses h, h' et h'', $\hat{\mathcal{L}}_T(h,h') \leq \hat{\mathcal{L}}_T(h,h'') + \hat{\mathcal{L}}_T(h',h'')$, on obtient :

$$\begin{aligned}
\hat{\mathcal{L}}_T(H_N,.) &\leq \hat{\mathcal{L}}_T(H_*,.) + \hat{\mathcal{L}}_T(H_N,H_*) \Leftrightarrow \\
\hat{\mathcal{L}}_T(H_N) &\leq \hat{\mathcal{L}}_T(H_*) + \hat{\mathcal{L}}_T(H_N,H_*) \\
&\leq \hat{\mathcal{L}}_T(H_*) + \hat{\mathcal{L}}_S(H_N,H_*) + |\hat{\mathcal{L}}_T(H_N,H_*) - \hat{\mathcal{L}}_S(H_N,H_*)| \\
&\leq \hat{\mathcal{L}}_T(H_*) + \hat{\mathcal{L}}_S(H_N,H_*) + \frac{1}{2}\hat{d}_{\mathcal{H}_\gamma}(S,T) \\
&\leq \hat{\mathcal{L}}_T(H_*) + \hat{\mathcal{L}}_S(H_N) + \hat{\mathcal{L}}_S(H_*) + \frac{1}{2}\hat{d}_{\mathcal{H}_\gamma}(S,T) \\
&= \hat{\mathcal{L}}_S(H_N) + \frac{1}{2}\hat{d}_{\mathcal{H}_\gamma}(S,T) + \lambda^*, \qquad (12)
\end{aligned}$$

où $\lambda^* = \hat{\mathcal{L}}_T(H_*) + \hat{\mathcal{L}}_S(H_*)$, que nous supposons faible quand l'adaptation est possible (Ben-David *et al.*, 2010). Le fait d'insérer l'Equation(12) dans l'Equation(10) permet de terminer la preuve. □

4.5. Critère d'arrêt de DABOOST

Le Théorème 5 signifie que la baisse de l'erreur en généralisation $\epsilon_{H_N^T}$ se fait en diminuant à la fois la perte empirique $\hat{\mathcal{L}}_S(H_N)$ sur S, ce qui est garanti par les propriétés de ADABOOST, mais aussi en réduisant la divergence empirique $\hat{d}_{\mathcal{H}_\gamma}(S,T)$. Cependant, le calcul de ce second terme requiert la connaissance des véritables étiquettes des exemples cible. Pour contourner ce problème, nous adaptons une stratégie présentée dans Ben-David *et al.* (2010) à notre contexte de marge en apprenant une hypothèse ayant des désaccords de marge (en fonction de γ) entre les exemples source et cible, ce qui ne nécessite pas de connaître les étiquettes des données. Comme DABOOST change d'espace de représentation (il construit des hyperplans dans un espace à N dimensions) afin de minimiser les deux différentes pertes ($\hat{L}_{H_N^T}$ et $\hat{\epsilon}_{H_N^S}$) en fonction de deux différents paramètres optimaux (α_n et β_n), nous proposons de regrouper à la fois les points source et cible dans le même espace, en les projetant dans l'espace à une dimension donné par $f(x) = \sum_n (\alpha_n + \beta_n) h_n(x)$. Ensuite, $\hat{d}_{\mathcal{H}_\gamma}(S,T)$ est calculée dans cet espace en évaluant la proportion de désaccords de marge entre S et T.

En considérant que d_h, $|S|$, γ et δ sont constants dans le Théorème 5 et que λ_* est supposé faible, trouver la meilleur hypothèse H_N, ou en d'autres termes déterminer le nombre optimal d'itérations N de DABOOST, peut être réalisé en gardant la combinaison linéaire de N^* hypothèses telle que

$$N^* = argmin_N \left(\hat{\mathcal{L}}_S(H_N) + \frac{1}{2}\hat{d}_{\mathcal{H}_\gamma}(S,T) \right). \qquad (13)$$

Donc, l'Equation(13) nous donne un critère d'arrêt théoriquement fondé pour DABOOST. En comparaison, un critère impliquant la $\mathcal{H}$-divergence serait :

$$N^* = argmin_N \left(\hat{\epsilon}_S(H_N) + \frac{1}{2}\hat{d}_{\mathcal{H}}(S,T) \right). \quad (14)$$

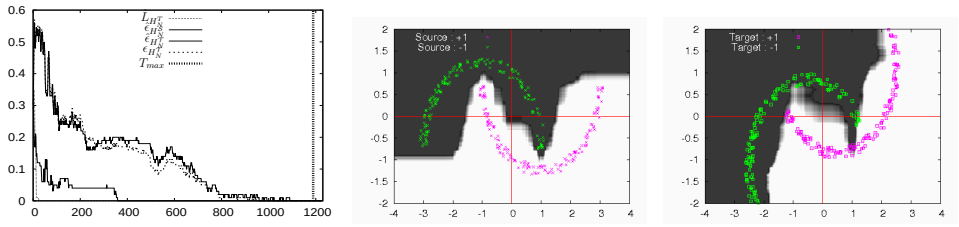
Nous comparons ces deux critères dans la section expérimentale ci-après.

5. Etude expérimentale

5.1. Base de données et protocole expérimental

Nous proposons d'évaluer notre approche sur un problème binaire "jouet" déjà utilisé pour l'évaluation de méthodes d'AD (e.g. Bruzzone & Marconcini (2010)). Nous considérons comme ensemble d'apprentissage deux lunes imbriquées dans un espace à 2 dimensions où les données sont issues d'une distribution uniforme pour chacune des deux lunes, correspondant aux deux classes (voir Figure 3(b)). Le domaine cible est obtenu en effectuant une rotation en sens anti-horaire sur le domaine source (Figure 3(c)). Nous considérons 4 problèmes de difficulté croissante en fonction de 4 angles de rotation, de 20 degrés à 50 degrés. Pour chaque domaine, nous générons 300 exemples (150 de chaque classe). Afin de prouver la capacité de généralisation de notre algorithme, nous évaluons les modèles inférés sur un ensemble de test indépendant composé de 1000 points distribués selon le domaine cible. Chaque problème d'adaptation est répété 10 fois et nous considérons le taux moyen d'erreur en classification obtenu sur l'échantillon test, en omettant le meilleur et le moins bon résultat. On fixe le nombre d'itérations de DABOOST à 1500 et on sélectionne la combinaison linéaire correspondant aux critères d'arrêt définis dans les Equations (13) et (14). Nous comparons notre approche avec ADABOOST et un classifieur SVM (avec un noyau Gaussien et des hyperparamètres tunés par validation croisée) appris uniquement sur la source. Nous la comparons également avec DASVM (en utilisant une implémentation LibSVM ainsi qu'une validation circulaire pour les hyperparamètres).

Par la suite, nous analysons dans un premier temps en profondeur le comportement de notre algorithme sur un exemple particulier avec pour objectif d'illustrer les résultats théoriques de la section 4. Nous présentons ensuite une étude comparative sur tous les problèmes.



(a) Comportement sur une tâche à 20 degrés. (b) Frontière de décision pour H_N^S sur la source. (c) Frontière de décision pour H_N^T sur la cible.

FIGURE 3: Illustration du comportement de DABOOST

5.2. Confirmation empirique des résultats théoriques

Nous présentons donc le comportement de DABOOST sur un problème de rotation à 20 degrés. En observant la Figure 3(a), nous pouvons faire les remarques suivantes : premièrement, comme attendu par rapport aux Théorèmes 2 et 3, les deux pertes empiriques convergent à 0 avec le nombre N d'itérations. Comme DABOOST concentre ses efforts sur l'adaptation au domaine cible, la convergence de la perte sur la cible $\hat{L}_{H_N^T}$ est atteinte très rapidement (après 25 itérations seulement). A cause de cette contrainte d'adaptation à la cible, la convergence de la perte sur la source $\hat{\epsilon}_{H_N^S}$ requiert quant à elle plus d'itérations (environ 380) mais permet aux résultats théoriques d'ADABOOST d'être toujours valables. De plus, l'erreur empirique de classification sur la cible $\hat{\epsilon}_{H_N^T}$, ainsi que celle sur l'ensemble de test $\epsilon_{H_N^T}$, décroît avec N et continue de baisser même quand les deux pertes empiriques ont atteint 0. Ce comportement confirme l'intérêt d'avoir une erreur faible sur la source ainsi que des marges importantes sur la cible. Finalement, nous pouvons observer que notre critère d'arrêt qui consiste à sélectionner la combinaison de classifieurs minimisant la borne empirique du Théorème 5 nous permet de choisir un classifieur final très efficace (ligne T_{max}).

Sur les Figures 3(b) et 3(c) sont représentées les zones de décision des modèles inférés. Les points sont étiquetés négatifs dans la région sombre, positifs dans celle plus claire, tandis que la région grise représente une zone d'indécision (qui ne correspond pas forcément à des erreurs). En observant la Figure 3(b), on peut remarquer que la frontière de décision apprise sur le domaine source (*i.e.* $H_N^S = \sum_n \alpha_n signe(h_n(\cdot))$) classifie correctement l'intégralité des exemples de l'ensemble d'apprentissage. Sur la Figure 3(c), nous avons reporté la frontière de décision apprise par DABOOST sur l'ensemble

Angle	20°	30°	40°	50°
SVM	10.32±0.78	24.01±0.92	32.16±0.85	40±1.08
DASVM	**0.00±0.00**	21.41±4.56	28.37±4.16	33.41±4.01
AdaBoost	29.26±0.50	32.05±1.57	44.32±0.10	53.78±0.00
DABoost-$\mathcal{H}$	4.01±1.49	4.76±0.93	13.00±6.19	28.65±11.17
DABoost	3.92±1.42	**4.41±0.46**	**11.57±1.79**	**22.67±7.77**
DABoost*	*0.41±0.26*	*2.32±0.44*	*4.83±0.68*	*9.37±2.37*

TABLE 1: Taux d'erreur et écarts types sur les problèmes des lunes.

cible ($H_N^T = \sum_n \beta_n signe(h_n(\cdot))$). On peut voir que la rotation a été presque parfaitement apprise. Les erreurs restantes correspondent aux exemples positifs situés en bas à droite de la région grise. Cette zone contient également des exemples négatifs du domaine source, ce qui explique pourquoi ces exemples (seulement trois d'entre eux) sont difficiles à apprendre. Nous rappelons ici que cette frontière de décision a été trouvée **sans aucune information sur les étiquettes des données cible**. Les frontières de décision des deux figures montrent bien l'intérêt de deux schémas de pondération différents.

5.3. Résultats généraux

Les résultats obtenus sur les différents problèmes d'adaptation (avec $\gamma = 0.2$) sont reportés dans la Table 1. Pour DABOOST, nous fournissons trois différents résultats : (i) **DABoost** qui utilise le critère d'arrêt basé sur notre H_γ-divergence, (ii) **DABoost**-$\mathcal{H}$ utilisant le critère d'arrêt basé sur la $\mathcal{H}$-divergence et **DABoost*** correspondant au meilleur modèle trouvé parmi toutes les itérations. Même si ce dernier résultat n'est pas justifié, il nous permet de connaître la meilleure performance possible réalisée par DABOOST. Nous pouvons faire les remarques suivantes : premièrement, excepté pour la rotation à 20 degrés (pour laquelle DASVM est légèrement meilleur), DABOOST réalise la meilleure performance de manière significative. Deuxièmement, notre critère d'arrêt est capable de sélectionner de meilleurs modèles que ceux choisis avec le critère basé sur la $\mathcal{H}$-divergence. Ceci peut être expliqué par la qualité de l'estimation de notre critère. En effet, en recherchant de grandes marges, nous prenons en compte la confiance des modèles, tandis que l'utilisation de la $\mathcal{H}$-divergence ne donne pas vraiment d'indication sur la qualité du modèle sur la cible. Finalement, on remarquera que DABOOST fournit des modèles encore meilleurs parmi toutes les itérations, ceci attestant du fait que notre critère d'arrêt pourrait être amélioré.

6. Conclusion

Dans ce papier, nous avons présenté un algorithme d'AD théoriquement fondé. Malgré sa simplicité algorithmique, DABOOST nous permet de dériver plusieurs garanties théoriques qui ont été confirmées sur une base de données artificielle. Si notre critère d'arrêt permet de trouver un bon modèle, nous avons noté qu'il existait une meilleure solution parmi toutes les itérations de DABOOST, ce qui ouvre la porte à de futurs travaux. Nous souhaitons également prouver les résultats de convergence de $\hat{d}_{\mathcal{H}_\gamma}$.

Références

BEN-DAVID S., BLITZER J., CRAMMER K., KULESZA A., PEREIRA F. & VAUGHAN J. W. (2010). A theory of learning from different domains. *Machine Learning*, **79**(1-2), 151–175.

BLITZER J., DREDZE M. & PEREIRA F. (2007). Biographies, bollywood, boomboxes and blenders : Domain adaptation for sentiment classification. In *ACL*.

BRUZZONE L. & MARCONCINI M. (2010). Domain adaptation problems : a DASVM classification technique and a circular validation strategy. *IEEE transactions on pattern analysis and machine intelligence*, **32**(5), 770–787.

CHELBA C. & ACERO A. (2006). Adaptation of maximum entropy capitalizer : Little data can help a lot. *Computer Speech & Language*, **20**(4), 382–399.

DAI W., YANG Q., XUE G.-R. & YU Y. (2007). Boosting for transfer learning. In *ICML*, p. 193–200.

FREUND Y. & SCHAPIRE R. E. (1996). Experiments with a new boosting algorithm. In *ICML*, p. 148–156.

MANSOUR Y., MOHRI M. & ROSTAMIZADEH A. (2008). Domain adaptation with multiple sources. In *NIPS*, p. 1041–1048.

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

MARTÍNEZ A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**(6), 748–763.

ROARK B. & BACCHIANI M. (2003). Supervised and unsupervised pcfg adaptation to novel domains. In *HLT-NAACL*.

SCHAPIRE R. E., FREUND Y., BARLETT P. & LEE W. S. (1997). Boosting the margin : A new explanation for the effectiveness of voting methods. In *ICML*.

VALIANT L. G. (1984). A theory of the learnable. *Commun. ACM*, **27**(11).