Adaptation de domaine de vote de majorité par auto-étiquetage non itératif

Emilie Morvant

Adaptation de domaine de vote de majorité par auto-étiquetage non itératif

Emilie Morvant

Institute of Science and Technology (IST) Austria, 3400 Klosterneuburg, Austria

13 juin 2014

Résumé

En apprentissage automatique, nous parlons d'adaptation de domaine lorsque les données de test (cibles) et d'apprentissage (sources) sont générées selon différentes distributions. Nous devons donc développer des algorithmes de classification capables de s'adapter à une nouvelle distribution, pour laquelle aucune information sur les étiquettes n'est disponible. Nous attaquons cette problématique sous l'angle de l'approche PAC-Bayésienne qui se focalise sur l'apprentissage de modèles définis comme des votes de majorité sur un ensemble de fonctions. Dans ce contexte, nous introduisons PV-MinCq une version adaptative d'un algorithme d'apprentissage de vote : MinCq. PV-MinCq suit le principe suivant. Nous transférons les étiquettes sources aux points cibles proches pour ensuite appliquer MinCq sur l'échantillon cible "auto-étiqueté" (justifié par une borne théorique). Plus précisément, nous définissons un auto-étiquetage non itératif qui se focalise dans les régions où les distributions marginales source et cible sont les plus similaires. Dans un second temps, nous étudions l'influence de notre auto-étiquetage pour en déduire une procédure de validation des hyperparamètres. Finalement, notre approche montre des résultats empiriques prometteurs.

Mots-clés : Adaptation de domaine, Vote de majorité, Théorie PAC-Bayésienne

1 Introduction

En apprentissage automatique, un des cadres théoriques les plus communs suppose l'échantillon d'apprentissage représentatif des données que l'on désire classer. Bien que cette hypothèse soit parfois une bonne approximation de la réalité, elle reste difficile à vérifier en pratique. Prenons par exemple un système de filtrage de spams : un système performant pour un utilisateur donné ne le sera pas nécessairement pour un utilisateur recevant des e-mails de natures différentes. Il faut alors être capable d'adapter le système d'un utilisateur à un autre. Ce problème peut être modélisé d'un point de vue statistique : l'échantillon d'apprentissage n'est alors plus issu de la même distribution de probabilité que les données à traiter, on parle d'adaptation de domaine [1]. De nos jours, cette problématique s'avère très active en apprentissage automatique, mais aussi dans de nombreuses communautés scientifiques telles qu'en multimédia, en traitement d'images, en traitement automatique de la langue, en bio-informatique. En effet, entre la grande diversité des données accessibles par Internet et le fait que la personnalisation soit au cœur de beaucoup de problématiques, toutes ces communautés s'intéressent à tirer au mieux parti de toutes les informations disponibles afin d'adapter ou de transférer les connaissances dont on dispose sur de nouveaux types de données. Plus formellement, à partir de données sources étiquetées, l'adaptation de domaine a pour objectif d'apprendre un modèle performant sur des données cibles pour lesquelles nous ne disposons d'aucune étiquette (ou de quelques étiquettes) [2]. Dans ce papier, nous proposons un nouvel algorithme d'adaptation de domaine sans étiquette cible, un scénario connu pour être plus difficile à traiter [BDU12].

Pour répondre à cette problématique, différents

1. *Domain adaptation* en anglais, voir [Mar11] pour un état de l'art. Notons que l'adaptation de domaine est un des champs d'étude de l'apprentissage par transfert [PY10, QCSSL09].

2. La tâche sans étiquette cible est souvent appelée adaptation de domaine non supervisée, celle avec quelques étiquettes cibles est l'adaptation de domaine semi-supervisée.

principes algorithmiques ont été proposés dans la littérature. Nous pouvons tout d'abord citer les méthodes de repondération de l'échantillon étiqueté pour qu'il "ressemble" le plus possible à l'échantillon cible au sens de la fonction de perte considérée. Cette approche permet de s'attaquer à des problématiques telles que le *covariate-shift*, où les distributions ne diffèrent que par leur marginale (*e.g.* [HSG$^+$07]). Une autre technique consiste à exploiter des procédures d'auto-étiquetage des données. Cependant, les algorithmes proposés sont souvent itératifs et lourds à mettre en œuvre (*e.g.* [HPS13, BM10]). L'approche DASVM proposée par [BM10] est l'une des méthodes de référence, son principe est le suivant. À chaque itération, DASVM apprend un classifieur SVM à partir des exemples sources étiquetés, puis certains sont remplacés par des données cibles auto-étiquetées via ce classifieur SVM [3]. Une troisième approche tire avantage d'une mesure de divergence entre distributions en suivant l'intuition que minimiser cette dernière, tout en gardant de bonnes performances sur les données sources, permet de quantifier plus "facilement" les garanties en généralisation. Différents travaux ont proposé de telles mesures et ont amené à des analyses de l'adaptation de domaine. Les divergences les plus utilisées, telles que la $\mathcal{H}\Delta\mathcal{H}$-divergence [BBCP07, BBC$^+$10] et la *discrepancy* [MMR08], mettent en jeu le désaccord entre les classifieurs possibles. Bien qu'elles produisent des analyses différentes, l'idée sous-jacente reste la même : le désaccord entre classifieurs doit être contrôlé tout en veillant à ce que les performances sur les données sources restent correctes. Bien évidemment, d'autres divergences existent pour quantifier la différence entre deux distributions et pourraient être étudiées dans un objectif adaptatif. Nous pouvons, par exemple, citer la *perturbed variation* [HM12] pour laquelle deux échantillons sont similaires si chaque point d'un échantillon est proche d'un point de l'autre échantillon. Nous allons utiliser cette mesure pour proposer une méthode non itérative d'auto-étiquetage des données cibles.

Dans ce papier, nous considérons la problématique particulière de l'adaptation de domaine PAC-Bayésienne (introduite dans [GHLM13]) qui se focalise sur l'apprentissage d'un modèle cible ayant la forme d'un vote de majorité pondéré sur un ensemble de classifieurs (ou votants). Leur analyse se base sur une divergence entre distributions. Cette dernière, appelée le ρ-désaccord, a été justifiée par une majoration précise de l'erreur du vote de majorité—la C-borne [LLM$^+$07]—et a l'avantage de considérer le désaccord entre les votants. Bien que leur étude théorique soit élégante, l'algorithme qui en dérive est restreint aux classifieurs linéaires et ne minimise pas directement l'erreur du vote. Notre objectif est donc de développer une méthode capable d'apprendre un vote sur un ensemble de fonctions à valeurs réelles dans le scénario de l'adaptation de domaine PAC-Bayésienne. Avec cet objectif en tête et sachant que la C-borne a amené à un algorithme simple et performant pour la classification supervisée, appelé MinCq [LMR11], nous proposons de le généraliser à l'adaptation de domaine grâce à un auto-étiquetage non itératif. Tout d'abord, nous formulons une nouvelle version de la C-borne appropriée à toute fonction d'(auto-)étiquetage (qui associe une étiquette à un exemple). Concrètement, nous définissons un auto-étiquetage qui se base sur la *perturbed variation* afin qu'il se concentre sur les régions dans lesquelles les marginales source et cible sont proches. Nous étiquetons ensuite l'échantillon cible uniquement dans ces régions (voir Figure 1, Section 3.2). Enfin, MinCq est appliqué sur cet échantillon auto-étiqueté. Outre ce nouvel algorithme d'adaptation, nous étudions l'influence de notre auto-étiquetage et en déduisons une procédure originale de sélection des hyperparamètres. Finalement, notre méthode globale, appelée PV-MinCq, montre des performances très prometteuses sur un jeu de données synthétique, meilleures qu'un auto-étiquetage défini à partir des k plus proches voisins et que d'autres méthodes d'adaptation de domaine.

Le papier est organisé comme suit. La Section 2 rappelle l'analyse PAC-Bayésienne de l'adaptation de domaine de [GHLM13], puis l'algorithme MinCq et ses bases théoriques en apprentissage supervisé [LMR11]. Dans la Section 3, nous présentons PV-MinCq, notre MinCq adaptatif basé sur un auto-étiquetage défini à l'aide de la *perturbed variation*. Avant de conclure, nous expérimentons notre méthode dans la Section 4.

2 Notations et contexte

Dans cette section, nous énonçons tout d'abord le cadre classique de l'approche PAC-Bayésienne en apprentissage supervisé. Puis, nous présentons l'analyse PAC-Bayésienne de l'adaptation de domaine [GHLM13], ainsi que l'algorithme d'apprentissage supervisé de vote de majorité : MinCq [LMR11].

3. Dans DASVM, les points auto-étiquetés sont ceux de plus faible confiance, et les points sources supprimés sont ceux de plus grande confiance.

2.1 Apprentissage supervisé et approche PAC-Bayésienne

Nous rappelons le cadre classique de la théorie PAC-Bayésienne—introduite dans [McA99]—offrant des bornes en généralisation (et des algorithmes) pour les votes de majorité pondérés sur un ensemble de fonctions à valeurs réelles, appelées votants.

Soit $X \subseteq \mathbb{R}^d$ l'espace de description des données de dimension d et $Y = \{-1, +1\}$ l'ensemble des étiquettes possibles. Un domaine est une distribution P_S sur $X \times Y$. La distribution d'un échantillon de taille m_s est notée $(P_S)^{m_s} = \bigotimes_{s=1}^{m_s} P_S$. La distribution marginale selon X de P_S est notée D_S. Nous considérons un échantillon d'apprentissage $S = \{(\mathbf{x}_s, y_s)\}_{s=1}^{m_s}$ composé de m_s exemples indépendemment et identiquement distribués (*i.i.d.*) selon $(P_S)^{m_s}$. Soit $\mathcal{H}$ un ensemble de votants à valeurs réelles tel que : $\forall h \in \mathcal{H},\ h : X \to \mathbb{R}$. Étant donné $\mathcal{H}$, les ingrédients de l'approche PAC-Bayésienne sont une distribution *a priori* π sur $\mathcal{H}$ (appelée le prior), un échantillon d'apprentissage S et une distribution *a posteriori* ρ sur $\mathcal{H}$ (appelée le posterior). Le prior π modélise la connaissance, avant l'observation de S, sur $\mathcal{H}$: les votants supposés meilleurs auront un plus grand poids selon π. Ensuite, étant donnée l'information portée par S, l'apprenant doit trouver le posterior ρ impliquant un vote de majorité pondéré B_ρ sur $\mathcal{H}$ avec de bonnes garanties en généralisation. Le vote B_ρ et ses risques réel et empirique sont définis comme suit.

Définition 1. *Soit $\mathcal{H}$ un ensemble de votants à valeurs réelles. Soit ρ une distribution sur $\mathcal{H}$. Le vote de majorité pondéré B_ρ est défini par :*

$$\forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

Le risque réel de B_ρ sur un domaine P_S et son risque empirique[4] *sur un échantillon S sont respectivement :*

$$\mathbf{R}_{P_S}(B_\rho) = \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P} y_s B_\rho(\mathbf{x}_s)\right),$$

$$\mathbf{R}_S(B_\rho) = \frac{1}{2}\left(1 - \frac{1}{m_s}\sum_{s=1}^{m_s} y_s B_\rho(\mathbf{x}_s)\right).$$

Les analyses PAC-Bayésienne classiques[5] ne se focalisent pas directement sur le risque de B_ρ, mais bornent le risque du classifieur stochastique de Gibbs G_ρ associé à ρ. Ce dernier étiquette un exemple $\mathbf{x}$ en tirant aléatoirement selon ρ un votant h dans $\mathcal{H}$, puis en retournant $h(\mathbf{x})$. Le risque de G_ρ correspond alors au moyennage des risques des votants de $\mathcal{H}$ selon ρ :

$$\mathbf{R}_P(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_P(h) = \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P} \mathop{\mathbf{E}}_{h\sim\rho} y_s h(\mathbf{x}_s)\right). \quad (1)$$

Signalons qu'il est admis dans la théorie PAC-Bayésienne que le classifieur déterministe B_ρ et le classifieur stochastique G_ρ sont reliés par :

$$\mathbf{R}_P(B_\rho) \leq 2\,\mathbf{R}_P(G_\rho). \quad (2)$$

2.2 L'étude PAC-Bayésienne de l'adaptation de domaine

Nous considérons maintenant le cadre de l'adaptation de domaine PAC-Bayésienne introduit dans [GHLM13]. La principale différence entre l'apprentissage supervisé et l'adaptation de domaine réside dans le fait que nous disposons de deux domaines différents sur $X \times Y$: le domaine source P_S et le domaine cible P_T (D_S et D_T étant les distributions marginales selon X respectives). L'objectif est alors d'apprendre un modèle performant sur le domaine cible P_T sachant que les seules étiquettes disponibles sont issues du domaine source P_S. Concrètement, dans le cadre décrit par [GHLM13], $S = \{(\mathbf{x}_s, y_s)\}_{t=1}^{m_s} \sim (P_S)^{m_s}$ est l'échantillon source étiqueté et $T = \{\mathbf{x}_t\}_{t=1}^{m_t} \sim (D_T)^{m_t}$ est l'échantillon cible non étiqueté. Nous désirons donc apprendre à partir de S et T un vote de majorité pondéré avec le plus faible risque réel possible sur le domaine cible, c'est-à-dire avec de bonnes garanties en généralisation sur P_T. En rappelant que les bornes en généralisation PAC-Bayésienne étudient le risque du classifieur de Gibbs, les auteurs ont proposé une analyse sur son risque cible $\mathbf{R}_{P_T}(G_\rho)$. Leur résultat principal prend la forme du théorème suivant.

Théorème 1 (Généralisation du Th. 4 de [GHLM13] à des votants réels)**.** *Soit $\mathcal{H}$ un ensemble de votants à valeurs réelles. Pour toute distribution ρ sur $\mathcal{H}$, on a :*

$$\mathbf{R}_{P_T}(G_\rho) \leq \mathbf{R}_{P_S}(G_\rho) + \operatorname{dis}_\rho(D_S, D_T) + \lambda_\rho, \quad (3)$$

où $\operatorname{dis}_\rho(D_S, D_T)$ est le ρ-désaccord entre les marginales D_S et D_T et est défini par :

$$\operatorname{dis}_\rho(D_S, D_T) = \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left(\mathop{\mathbf{E}}_{\mathbf{x}_t\sim D_T} h(\mathbf{x}_t)h'(\mathbf{x}_t) - \mathop{\mathbf{E}}_{\mathbf{x}_s\sim D_S} h(\mathbf{x}_s)h'(\mathbf{x}_s)\right)\right|. \quad (4)$$

λ_ρ est un terme lié[6] *aux vrais étiquetages de P_S et P_T.*

4. Nous faisons appel à la fonction de perte linéaire puisque nos votants sont à valeurs réelles, mais dans le cas particulier de B_ρ elle est équivalente à la fonction de perte $0-1$.

5. Les analyses PAC-Bayésiennes classiques peuvent être trouvées dans [McA03, See02, Lan05, Cat07, GLLM09].

6. En pratique, nous ne pouvons pas calculer/estimer λ_ρ puisqu'il dépend des étiquettes cibles inconnues. Il est ainsi généralement négligé. Nous ne développons donc pas ce point dans ce papier, mais de plus amples détails peuvent être trouvés dans [GHLM13].

Il est important de remarquer que cette borne reflète de la philosophie classique en adaptation de domaine : il est admis qu'il est plus facile d'adapter un modèle à un domaine cible lorsqu'à la fois les domaines sont proches et qu'il existe un modèle performant sur le domaine source [BBCP07, BBC$^+$10, MMR08].

Les auteurs de [GHLM13] ont dérivé de cette analyse un premier algorithme appelé PBDA pour minimiser ce compromis entre risque source et ρ-désaccord. Bien que PBDA ait démontré l'utilité de l'approche PAC-Bayésienne en adaptation de domaine, il reste spécifique aux classifieurs linéaires, ne se focalise pas directement sur le vote de majorité appris B_ρ et n'améliore pas significativement les résultats empiriques de méthodes de l'état de l'art.

Dans ce papier, notre objectif est de s'attaquer à ces inconvénients en proposant un nouvel algorithme pour apprendre un vote de majorité adaptatif sur un ensemble de votants à valeurs réelles, en minimisant un risque sur ce vote. Pour ce faire, nous nous concentrons sur le ρ-désaccord défini dans l'Équation (4). En effet, ce terme trouve sa source dans la borne théorique (la C-borne [LLM$^+$07]) sur le risque source du vote $\mathbf{R}_{P_S}(B_\rho)$; borne ayant donné lieu à l'algorithme (non adaptatif) MinCq pour apprendre un vote de majorité sur un ensemble de votants à valeurs réelles [LMR11]. Nous rappelons maintenant ces résultats en apprentissage supervisé, puis nous les généralisons à l'adaptation de domaine en Section 3.1.

2.3 MinCq : un algorithme d'apprentissage supervisé de vote de majorité

L'Équation (2) qui relie le classifieur stochastique G_ρ au vote de majorité B_ρ peut être très imprécise. Les auteurs de [LLM$^+$07] et de [LMR11] ont récemment proposé la C-borne, une relation plus précise énoncée dans le Théorème 2 qui suit. Ce résultat se base sur la notion suivante de ρ-marge.

Définition 2 ([LMR11]). *La ρ-marge d'un exemple $(\mathbf{x}, y) \in X \times Y$ réalisée sur la distribution ρ sur $\mathcal{H}$ est donnée par :* $\mathop{\mathbf{E}}_{h\sim\rho} yh(\mathbf{x})$.

D'après la définition de B_ρ, il est trivial de montrer que B_ρ classe correctement un exemple $\mathbf{x}_s$ lorsque la ρ-marge est strictement positive. Ainsi, sous la convention que si $y_s\mathbf{E}_{h\sim\rho}\, h(\mathbf{x}_s) = 0$ alors B_ρ se trompe sur $(\mathbf{x}_s, y_s)$, pour tout domaine P_S sur $X \times Y$ on a :

$$\mathbf{R}_{P_S}(B_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x}_s,y_s)\sim P_S} \left(\mathop{\mathbf{E}}_{h\sim\rho} y_s h(\mathbf{x}_s) \leq 0 \right).$$

D'après cette égalité, l'inégalité de Cantelli-Chebitchev permet de démontrer la borne suivant sur $\mathbf{R}_P(B_\rho)$ [LLM$^+$07, LMR11].

Théorème 2 (La C-borne comme exprimée dans [LMR11]). *Pour toute distribution ρ sur $\mathcal{H}$ et pour tout domaine P_S sur $X \times Y$ de marginale (selon X) D_S, si* $\mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P_S} y_s h(\mathbf{x}_s) > 0$, *alors :*

$$\mathbf{R}_{P_S}(B_\rho) \leq 1 - \frac{\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{(\mathbf{x}_s,y_s)\sim P_S} y_s h(\mathbf{x}_s) \right)^2}{\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{\mathbf{x}_s\sim D_S} h(\mathbf{x}_s)h'(\mathbf{x}_s)}.$$

Le numérateur de cette borne correspond, en fait, au premier moment statistique de la ρ-marge de B_ρ réalisée sur P_S, qui peut être relié au risque du classifieur de Gibbs (voir Équation (1)). Le dénominateur, quant à lui, est le second moment statistique de cette ρ-marge, qui peut être vu comme une mesure de désaccord entre les paires de votants de $\mathcal{H}$ (plus cette valeurs est faible, plus les votants sont en désaccord) et peut être relié au ρ-désaccord (voir l'Équation (4)).

Dans le cadre de l'apprentissage supervisé, les auteurs de [LMR11] ont démontré une élégante borne en généralisation PAC-Bayésienne justifiant de la minimisation empirique de la C-borne pour apprendre un vote de majorité sur $\mathcal{H}$. Ils en ont déduit un algorithme quadratique simple appelé MinCq et décrit dans l'Algorithme 1. Concrètement, MinCq apprend un vote de majorité en optimisant la C-borne empirique mesurée sur l'échantillon d'apprentissage S : MinCq minimise le dénominateur, c'est-à-dire maximise le désaccord (Équation (5)), sous la contrainte que le numérateur soit fixe, c'est-à-dire que le risque du classifieur de Gibbs soit fixe (Équation (6)), et sous une régularisation particulière (Équation (7)) [7]. Signalons que MinCq a démontré de bonnes performances sur des tâches de classification supervisée.

Avec le point de vue de l'adaptation de domaine, la C-borne et MinCq se focalisent sur le compromis suggéré par le Théorème 1. En effet, la définition du ρ-désaccord (Équation (4)) est étroitement reliée à la C-borne selon l'affirmation suivante : si les risques source et cible du classifieur de Gibbs sont similaires, alors les risques source et cible du vote sont similaires lorsque la différence entre les désaccords source et cible sont proches.

Par la suite, nous proposons de faire appel à la C-borne et à MinCq pour définir une méthode originale et générale d'apprentissage de vote de majorité

7. Les détails techniques sont énoncés dans [LMR11].

Algorithme 1 MinCq$(S, \mathcal{H}, \mu)$

entrée Un échantillon $S = \{(\mathbf{x}_s, y_s)\}_{s=1}^{m_s} \sim (P_S)^{m_s}$, n votants $\mathcal{H} = \{h_1, \ldots, h_n\}$, une marge désirée $\mu > 0$

sortie $B_\rho(\cdot) = \text{sign}\left[\sum_{j=1}^{n} \left(2\rho_j - \frac{1}{n}\right) h_j(\cdot)\right]$

$$\textbf{Résoudre} \quad \underset{\boldsymbol{\rho}}{\text{argmin}} \quad \boldsymbol{\rho}^T \mathbf{M} \boldsymbol{\rho} - \mathbf{A}^T \boldsymbol{\rho}, \tag{5}$$

$$\textbf{s.c.} \quad \mathbf{m}^T \boldsymbol{\rho} = \frac{\mu}{2} + \frac{1}{2nm_s} \sum_{j=1}^{n} \sum_{s=1}^{m_s} y_s h_j(\mathbf{x}_s), \tag{6}$$

$$\forall j \in \{1, \ldots, n\}, \quad 0 \leq \rho_j \leq \tfrac{1}{n}, \tag{7}$$

où $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)^\top$ est un vecteur de poids,
$\mathbf{M}$ est la matrice de dimension $n \times n$ telle que : $\forall (j, j') \in \{1, \ldots, n\}^2$, $\sum_{s=1}^{m_s} \frac{h_j(\mathbf{x}_s) h_{j'}(\mathbf{x}_s)}{m_s}$,

$$\mathbf{A} = \left(\sum_{j=1}^{n} \sum_{s=1}^{m_s} \frac{h_1(\mathbf{x}_s) h_j(\mathbf{x}_s)}{nm_s}, \ldots, \sum_{j=1}^{n} \sum_{s=1}^{m_s} \frac{h_n(\mathbf{x}_s) h_j(\mathbf{x}_s)}{nm_s} \right)^\top,$$

$$\text{et } \mathbf{m} = \left(\tfrac{1}{m_s} \sum_{s=1}^{m_s} y_s h_1(\mathbf{x}_s), \ldots, \tfrac{1}{m_s} \sum_{s=1}^{m_s} y_s h_n(\mathbf{x}_s) \right)^\top$$

sur un ensemble de votants à valeurs réelles dans le scénario de l'adaptation de domaine.

3 Un MinCq Adaptatif

Afin de tirer parti de l'algorithme MinCq, nous étendons tout d'abord la C-borne au cadre de l'adaptation de domaine.

3.1 Une C-borne pour l'adaptation de domaine avec auto-étiquetage

Soit une fonction d'étiquetage $l : X \to Y$, qui associe une étiquette $y \in Y$ à un exemple (cible) non étiqueté $\mathbf{x}_t \sim D_T$. Nous formulons la C-borne comme suit.

Corollaire 3. *Pour toute distribution ρ sur $\mathcal{H}$, pour tout domaine P_T sur $X \times Y$ de marginal (sur X) D_T, pour toute fonction d'étiquetage $l : X \to Y$ telle que $\mathop{\mathbf{E}}_{h \sim \rho} \mathop{\mathbf{E}}_{\mathbf{x}_t \sim D_T} l(\mathbf{x}_t) h(\mathbf{x}_t) > 0$, on a :*

$$\mathbf{R}_{P_T}(B_\rho) \leq 1 - \frac{\left(\mathop{\mathbf{E}}_{h \sim \rho} \mathop{\mathbf{E}}_{\mathbf{x} \sim D_T} l(\mathbf{x}_t) h(\mathbf{x}_t) \right)^2}{\mathop{\mathbf{E}}_{(h,h') \sim \rho^2} \mathop{\mathbf{E}}_{\mathbf{x}_t \sim D_T} h(\mathbf{x}_t) h'(\mathbf{x}_t)} + \frac{1}{2} \left| \mathop{\mathbf{E}}_{(\mathbf{x}_t, y_t) \sim P_T} \left(y_t - l(\mathbf{x}_t) \right) \right|.$$

Démonstration. Le résultat s'obtient grâce à l'égalité suivante :

$$\left| \mathbf{R}_{P_T}(B_\rho) - \mathbf{R}_{\widehat{P_T}}(B_\rho) \right| = \tfrac{1}{2} \left| \mathop{\mathbf{E}}_{(\mathbf{x}_t, y_t) \sim P_T} \left(y_t - l(\mathbf{x}_t) \right) \right|,$$

où : $\mathbf{R}_{\widehat{P_T}}(B_\rho) = \frac{1}{2} \Big(1 - \mathop{\mathbf{E}}_{\mathbf{x}_t \sim D_T} l(\mathbf{x}_t) B_\rho(\mathbf{x}_t) \Big)$. □

Nous reconnaissons la C-borne du Théorème 2 où la vraie étiquette y_t d'un exemple $\mathbf{x}_t$ a été remplacée par $l(\mathbf{x})$. Le terme $\frac{1}{2} \left| \mathbf{E}_{(\mathbf{x}_t, y_t) \sim P_T} \left(y_t - l(\mathbf{x}_t) \right) \right|$ peut alors être vu comme une divergence entre le vrai étiquetage et celui proposé par l, puisqu'il calcule l'écart entre l'étiquetage estimé et le vrai étiquetage : plus l et la vraie fonction d'étiquetage sont proches, plus la borne est précise. Notons que les bornes en généralisation proposées dans [LMR11] restent valides.

Il est important de remarquer que seul un domaine apparaît dans cette borne. Si nous supposons que ce dernier est le domaine cible, il est pertinent de définir une fonction d'étiquetage en s'aidant de l'information portée par l'échantillon étiqueté source S. Pour répondre à cette problématique de définition d'une fonction d'étiquetage, que nous appelons une fonction d'auto-étiquetage, nous suivons l'intuition suivante : étant donné un exemple source étiqueté $(\mathbf{x}_s, y_s) \in S$, nous désirons transférer son étiquette y_s à un point cible $\mathbf{x}_t$ proche de $\mathbf{x}_s$. Nous proposons donc d'explorer la *perturbed variation* [HM12], une récente mesure de divergence entre distributions basée sur cette intuition. Par la suite, nous définissons une fonction d'auto-étiquetage à l'aide de cette divergence. Cette dernière nous permet d'auto-étiqueter l'échantillon cible T, sur lequel nous pouvons appliquer MinCq (justifié par le Corollaire 3).

3.2 Un MinCq adaptatif via un auto-étiquetage non itératif

Avant de présenter notre auto-étiquetage, nous rappelons la définition de la *perturbed variation*.

Définition 3 ([HM12])**.** *Soit D_S et D_T deux distributions sur X et $M(D_S, D_T)$ l'ensemble des distributions jointes sur $X \times X$ de marginales D_S et D_T. La* perturbed variation *vis-à-vis d'une distance $d : X \times X \to \mathbb{R}^+$ et d'un rayon $\epsilon > 0$ est :*

$$PV(D_S, D_T, \epsilon, d) = \inf_{\nu \in M(D_S, D_T)} \mathop{\mathbf{Pr}}_{\nu} \left[d(\mathcal{X}, \mathcal{X}') > \epsilon \right],$$

sur toutes les paires $(D_S, D_T) \sim \nu$, telles que la marginale de $\mathcal{X}$ (resp. $\mathcal{X}'$) soit D_S (resp. D_T).

Algorithme 2 $\widehat{PV}(S,T,\epsilon,d)$

entrée $S=\{\mathbf{x}_s\}_{s=1}^{m_s}$, $T=\{\mathbf{x}_t\}_{t=1}^{m_t}$ des échantillons non étiquetés, un rayon $\epsilon>0$, une distance $d: X\times X\to\mathbb{R}^+$

sortie $\widehat{PV}(S,T,\epsilon,d)$

1. $G\leftarrow \Big(V=(A,B),E\Big)$ tel que :
   $A=\{\mathbf{x}_s\in S\}$, $B=\{\mathbf{x}_t\in T\}$, $e_{st}\in E$ si $d(\mathbf{x}_s,\mathbf{x}_t)\leq\epsilon$
2. $M_{ST}\leftarrow$ couplage maximal sur G
3. $(S_u,T_u)\leftarrow$ points de S, resp. de T, non couplés
4. Retourner $\widehat{PV}(S,T,\epsilon,d)=\frac{1}{2}\Big(\frac{|S_u|}{m_s}+\frac{|T_u|}{m_t}\Big)$

Algorithme 3 PV-MinCq$(S,T,\mathcal{H},\mu,\epsilon,d)$

entrée Un échantillon source $S=\{(\mathbf{x}_s,y_s)\}_{s=1}^{m_s}$ et un cible $T=\{\mathbf{x}_t\}_{t=1}^{m_t}$, un ensemble de votants $\mathcal{H}$, une marge désirée $\mu>0$, un rayon $\epsilon>0$, une distance $d: X\times X\to\mathbb{R}^+$

sortie $B_\rho(\cdot)$

1. $M_{ST}\leftarrow$ Étapes 1. et 2. de l'Algorithme 2
2. $\widehat{T}\leftarrow\{(\mathbf{x}_t,y_s):(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST},\mathbf{x}_t\in T,(\mathbf{x}_s,y_s)\in S\}$
3. Retourner MinCq$(\widehat{T},\mathcal{H},\mu)$

En d'autres termes, deux échantillons sont similaires si chaque point cible est proche d'un point source : on cherche la distribution jointe ν telle que la probabilité pour $(\mathcal{X},\mathcal{X}')\sim\nu$ de l'évènement "être à une distance plus élevée que ϵ" soit minimale. Notons que cette mesure est consistante et que son estimation empirique $\widehat{PV}(S,T,\epsilon,d)$ peut efficacement être calculée à l'aide d'une procédure de couplage maximal décrite dans l'Algorithme 2 [HM12].

Pour définir notre auto-étiquetage, nous faisons appel au couplage maximal M_{ST} obtenu à l'étape 2 de l'Algorithme 2. Plus précisément, nous étiquetons les exemples cibles de T grâce à M_{ST} comme suit : si $\mathbf{x}_t\in T$ appartient à un couple $(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST}$, alors $\mathbf{x}_t$ aura pour étiquette y_s de $\mathbf{x}_s$; sinon nous retirons $\mathbf{x}_t$ de T. L'échantillon auto-étiqueté $\widehat{T}$ construit est alors :

$$\widehat{T}=\{(\mathbf{x}_t,y_s):(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST},\mathbf{x}_t\in T,(\mathbf{x}_s,y_s)\in S\}.$$

En fait, nous restreignons l'adaptation aux régions où les marginales source et cible coïncident selon la mesure d. Nous appliquons ensuite MinCq sur $\widehat{T}$. Notre procédure d'auto-étiquetage est illustrée par la Figure 1 et notre algorithme global, appelé PV-MinCq, est présenté dans l'Algorithme 3.

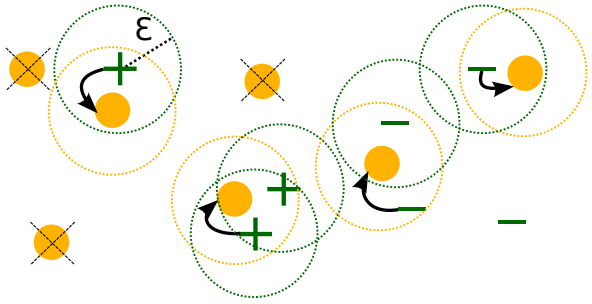

FIGURE 1 – Illustration de l'auto-étiquetage basé sur la *perturbed variation*. Les exemples sources étiquetés sont en vert (foncé), les exemples cibles non étiquetés en orange (clair). Les cercles représentent les candidats au couplage. Les flèches correspondent au points finalement couplés M_{ST} et donc à la fonction d'auto-étiquetage. Les exemples cibles non couplés sont supprimés. La quantité des points sources et cibles non couplés indiquent la valeur de la *perturbed variation*

3.3 Étude de l'auto-étiquetage

Dans cette section, nous discutons de l'impact de notre procédure d'auto-étiquetage et du choix de la distance d intervenant dans le calcul du couplage. Étant donnée une tâche d'adaptation de domaine, nous définissons tout d'abord la notion de "distance correcte".

Définition 4. *Étant donné un ensemble de votants $\mathcal{H}$ et un rayon $\epsilon>0$, une distance $d: X\times X\to\mathbb{R}^+$ est $\epsilon(\mathcal{H})$-correcte pour une tâche d'adaptation de domaine de P_S vers P_T, s'il existe $\epsilon(\mathcal{H})\geq 0$ tel que :*

$$\epsilon(\mathcal{H})=\max_{\substack{h\in\mathcal{H},\\(\mathbf{x}_t,\mathbf{x}_s)\sim D_S\times D_T,\\ d(\mathbf{x}_t,\mathbf{x}_s)\leq\epsilon}}\left|h(\mathbf{x}_s)-h(\mathbf{x}_t)\right|.$$

En d'autres termes, la propriété naturelle suivante doit être vérifiée : si $\mathbf{x}_t$ et $\mathbf{x}_s$ sont proches selon d, alors pour tous les votants de $\mathcal{H}$ l'écart entre les valeurs retournées $h(\mathbf{x}_s)$ et $h(\mathbf{x}_t)$ est faible. Étant donné $\mathcal{H}$, $\epsilon>0$ et une $\epsilon(\mathcal{H})$-correcte $d: X\times X\to\mathbb{R}^+$, on considère le couplage M_{ST} obtenu à l'étape 2 de l'Algorithme 2. Par définition, pour tout couple $(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST}$, $\mathbf{x}_t$ et $\mathbf{x}_s$ partage la même étiquette y_s et nous avons $d(\mathbf{x}_t,\mathbf{x}_s)\leq\epsilon$. Nous étudions maintenant l'influence de d et $\epsilon(\mathcal{H})$ sur la borne PAC-Bayésienne du Théorème 1 en se restreignant à M_{ST}. Nous avons besoin des notations suivantes. Les sous-échantillons source et cible

associés à M_{ST} sont respectivement :

$$\widehat{S} = \{(\mathbf{x}_s, y_s) : (\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}, \mathbf{x}_t \in T, (\mathbf{x}_s, y_s) \in S\},$$
$$\widehat{T} = \{(\mathbf{x}_t, y_s) : (\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}, \mathbf{x}_t \in T, (\mathbf{x}_s, y_s) \in S\}.$$

Premièrement, nous majorons l'écart entre les risques de G_ρ sur $\widehat{S}$ et $\widehat{T}$. Pour tout ρ sur $\mathcal{H}$, pour tout couple $(\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}$, on a :

$$\begin{aligned}
&\left| \tfrac{1}{2}\left(1 - y_s \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}_t)\right) - \tfrac{1}{2}\left(1 - y_s \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}_s)\right) \right| \\
&= \tfrac{1}{2} \left| \mathop{\mathbf{E}}_{h\sim\rho} (h(\mathbf{x}_t) - h(\mathbf{x}_s)) \right| \\
&\leq \tfrac{1}{2} \mathop{\mathbf{E}}_{h\sim\rho} |h(\mathbf{x}_t) - h(\mathbf{x}_s)| = \tfrac{1}{2} \mathop{\mathbf{E}}_{h\sim\rho} \epsilon(\mathcal{H}) = \tfrac{1}{2}\epsilon(\mathcal{H}).
\end{aligned}$$

Alors : $\left|\mathbf{R}_{\widehat{T}}(G_\rho) - \mathbf{R}_{\widehat{S}}(G_\rho)\right| \leq \frac{1}{2}\epsilon(\mathcal{H})$.
Ainsi, les risques empiriques du classifieur de Gibbs sur les échantillons source $\widehat{S}$ et cible $\widehat{T}$ diffèrent au plus de $\frac{1}{2}\epsilon(\mathcal{H})$: plus $\epsilon(\mathcal{H})$ est faible, plus les risques sont proches. Minimiser $\mathbf{R}_{\widehat{T}}(G_\rho)$ minimise alors $\mathbf{R}_{\widehat{S}}(G_\rho)$.

Deuxièmement, nous pouvons majorer l'écart entre le désaccord entre votants sur $\widehat{S}$ et $\widehat{T}$. Pour tout ρ sur $\mathcal{H}$ et pour tout $(\mathbf{x}_t, \mathbf{x}_s) \in M_{ST}$, on a :

$$\begin{aligned}
&\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} [h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t)] \right| \\
&\leq \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \Big[(\epsilon(\mathcal{H}) + h(\mathbf{x}_t))(\epsilon(\mathcal{H}) + h'(\mathbf{x}_t)) - h(\mathbf{x}_t)h'(\mathbf{x}_t)\Big] \right| \\
&= \left| \epsilon(\mathcal{H})^2 + 2 \mathop{\mathbf{E}}_{h\sim\rho} \epsilon(\mathcal{H}) h(\mathbf{x}_t) \right|
\end{aligned}$$

Alors, le ρ-désaccord entre $\widehat{S}$ et $\widehat{T}$ peut s'écrire :

$$\begin{aligned}
&\mathrm{dis}_\rho(\widehat{S}, \widehat{T}) \\
&= \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST}} [h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t)] \right| \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x}_t,\mathbf{x}_s)\in M_{ST}} \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} [h(\mathbf{x}_s)h'(\mathbf{x}_s) - h(\mathbf{x}_t)h'(\mathbf{x}_t)] \right| \\
&\leq \epsilon(\mathcal{H}) \left(1 + 2 \mathop{\mathbf{E}}_{(\mathbf{x}_t)\in\widehat{T}} \left| \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}_t) \right| \right)
\end{aligned}$$

Dans une telle situation, la divergence $\mathrm{dis}_\rho(\widehat{S}, \widehat{T})$ entre deux échantillons est bornée par un terme dépendant de $\epsilon(\mathcal{H})$ et de la confiance du vote de majorité sur $\widehat{T}$.

Ces résultats suggèrent que nous devons minimiser $\epsilon(\mathcal{H})$, tout en gardant de bonnes performances sur $\widehat{T}$. Cela légitimise notre méthode qui (i) transfère les étiquettes sources sur le domaine cible afin de rapprocher les risques source et cible du classifieur de Gibbs,

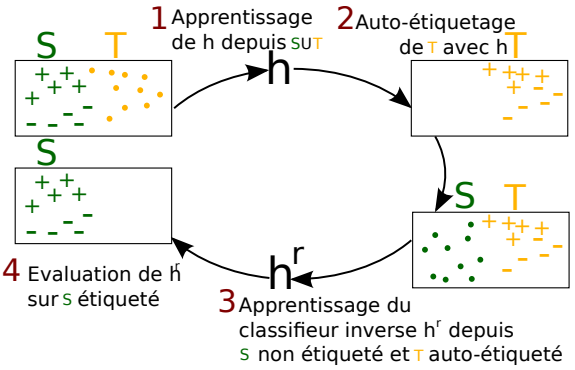

FIGURE 2 – Le principe de la validation inverse.

(ii) puis applique MinCq pour optimiser le désaccord entre votants sur l'échantillon cible (étant donné un risque de Gibbs sur les auto-étiquettes). Cependant, bien que l'on puisse choisir ϵ (et donc $\epsilon(\mathcal{H})$) aussi faible que désiré, un petit ϵ implique un ensemble de couples M_{ST} plus petit et donc une *perturbed variation* plus grande. Dans ce cas, la taille de $\widehat{T}$ tend à diminuer, impliquant que les garanties pour le classifieur de Gibbs diminuent. Afin d'éviter ce comportement, nous exploitons cette propriété dans la section suivante pour définir une technique de validation des hyperparamètres.

3.4 Validation des hyperparamètres

Une dernière question concerne la sélection des hyperparamètres μ et ϵ. Habituellement en adaptation de domaine, nous pouvons faire appel à une validation inverse ou circulaire [BM10, ZFY+10], avec l'idée que si les domaines sont étroitement reliés, alors un classifieur inverse, appris à partir des données cibles étiquetées par le modèle courant, montrera de bonnes performances sur les données sources (voir la Figure 2 pour l'intuition). Cependant, la première étape de PV-MinCq vise à transférer les étiquettes sources. Comme vu précédemment, notre objectif principal est alors de valider ce transfert. La validation inverse n'a donc ici plus vraiment de sens.

Nous proposons alors d'utiliser l'analyse précédente en faisant appel à toute l'information disponible, c'est-à-dire des échantillons originaux S et T. Nous avons montré que le ρ-désaccord peut être majoré par un terme dépendant de l'auto-étiquetage basé sur la *perturbed variation* (et de la confiance du vote sur cet étiquetage). Ainsi, la *perturbed variation* entre D_S et D_T doit être contrôlée : plus sa valeur est faible, plus les échantillons sont similaires. Cependant, minimiser

la *perturbed variation* en fonction de ϵ est très simple : il est possible de trouver une valeur élevée[8] de ϵ, impliquant une faible *perturbed variation*. Pour compenser ce comportement, nous pouvons alors contrôler la performance. En effet, plus ϵ est élevé, plus la distance entre les exemples source et cible d'un couple (de M_{ST}) est grande. Les points couplés sont donc plus éloignés au sens de d, ce qui va impliquer un écart plus grand entre les risques source et cible. Un tel comportement peut amener à une perte de performances sur l'échantillon source original. Ainsi, un auto-étiquetage pertinent correspond à celui optimisant le compromis :

$$\mathbf{R}_S(B_\rho) + \widehat{PV}(S, T, \epsilon, d),$$

où $\mathbf{R}_S(B_\rho)$ est le risque empirique sur l'échantillon source et $\widehat{PV}(S, T, \epsilon, d)$ est la *perturbed variation* entre S et T. Il est intéressant de constater que cette procédure est reliée à la philosophie de l'adaptation de domaine : nous souhaitons minimiser une divergence entre les domaines tout en gardant de bonnes performances sur le domaine source.

Concrètement, pour tout ensemble de paramètres possible (μ, ϵ) et étant donné k sous-échantillons de S ($S = \cup_{i=1}^{k} S_i$), PV-MinCq apprend un vote de majorité B_ρ à partir des $k-1$ échantillons étiquetés de S (et T). Puis, B_ρ est évalué sur le dernier $k^{\text{ème}}$ échantillon. Son risque empirique correspond alors à la moyenne des erreurs sur les k échantillons :

$$\mathbf{R}_S(B_\rho) = \frac{1}{k} \sum_{i=1}^{k} \mathbf{R}_{S_i}(B_\rho),$$

et $\widehat{PV}(S, T, \epsilon, d)$ est calculée par l'Algorithme 2.

4 Résultats expérimentaux

Dans cette section, nous évaluons notre méthode PV-MinCq pour apprendre un vote sur un ensemble de noyaux Gaussien définis à partir de l'échantillon source. Nous comparons PV-MinCq aux méthodes suivantes :

- SVM appris uniquement à partir de l'échantillon source (sans adaptation) ;
- MinCq [LMR11] appris uniquement à partir de l'échantillon source ;
- TSVM [Joa99], un SVM transductif semi-supervisé[9], appris à partir des deux domaines ;

8. Par exemple, si ϵ est égal à la plus grande distance entre un exemple source et un exemple cible.

9. TSVM n'est pas un algorithme d'adaptation à proprement dit, mais il implique, en général, des résultats très intéressants dans des scénarios adaptatifs.

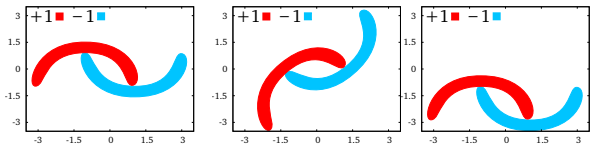

FIGURE 3 – À gauche : le domaine source. À droite : un domaine cible avec 40° de rotation, et le domaine cible translaté.

- DASVM [BM10], un algorithme adaptatif basé sur un auto-étiquetage itératif ;
- DASF [MHA12], un algorithme d'adaptation basé sur l'analyse de [BBCP07] et minimisant le compromis entre divergence et risque source ;
- PBDA [GHLM13], l'algorithme d'adaptation PAC-Bayésienne pour minimiser la borne du Théorème 1 ;
- PV-SVM, pour lequel on définit l'auto-étiquetage de l'échantillon cible comme pour PV-MinCq, puis on applique un SVM classique sur les auto-étiquettes ;
- NN-MinCq qui utilise un auto-étiquetage basé sur les k-PPV : nous étiquetons un point cible via un classifieur k-PPV dont les prototypes proviennent de l'échantillon source (k est tuné).

Pour calculer l'auto-étiquetage basé sur la *perturbed variation*, nous utilisons la distance euclidienne. Chaque paramètre est sélectionné à l'aide d'une grille de recherche (avec cinq sous-échantillons) via une validation croisée classique pour SVM, MinCq et TSVM, une validation inverse pour DASVM, DASF, PBDA et NN-MinCq, et la validation décrite dans la Section 3.4 pour PV-SVM et PV-MinCq.

Nous nous attaquons au problème de classification binaire appelé "lunes jumelles", où chaque lune correspond à une étiquette (voir la Figure 3). Nous considérons sept domaines cibles différents définis par sept rotations anti-horaires du domaine source (de 20° à 80°). Plus l'angle est élevé, plus l'adaptation est difficile. En outre, nous considérons un domaine cible défini par une translation du domaine source. Nous générons aléatoirement 150 exemples positifs et 150 négatifs pour chaque domaine. Afin d'estimer l'erreur en généralisation de notre approche, chaque algorithme est évalué sur un ensemble de test de 1 500 exemples cibles. Chaque tâche est répétée dix fois. Nous reportons les pourcentages moyens de classification correcte dans la Table 1. Nous faisons les remarques suivantes.

Tout d'abord, PV-MinCq produit en moyenne de meilleurs résultats que les autres approches, et apparaît plus robuste aux changements de densités (NN-MinCq

TABLE 1 – Pourcentages moyens de classification correcte sur dix tirages aléatoires pour les sept rotations, et pour la translation (*trans.*). Aucun résultat n'est reporté pour NN-MinCq pour la translation puisque, dans ce cas, la valeur de l'auto-étiquette est la même pour tous les exemples cibles.

Angle	20°	30°	40°	50°	60°	70°	80°	*trans.*
SVM	89.6	76	68.8	60	47.2	26.1	19.2	50.6
MinCq	92.1	78.2	69.8	61	50.1	40.7	32.7	50.7
TSVM	**100**	78.9	74.6	70.9	64.7	21.3	18.9	94.9
DASVM	**100**	78.4	71.6	66.6	61.6	25.3	21.1	50.1
PBDA	90.6	89.7	77.5	58.8	42.4	37.4	39.6	85.9
DASF	98.3	92.1	83.9	70.2	54.7	43	38.9	82.8
PV-SVM	94.2	82.5	75.1	67.7	55.2	43.6	30.3	97.1
NN-MinCq	97.7	83.7	77.7	69.2	58.1	47.9	42.1	∅
PV-MinCq	99.9	**99.7**	**99**	**91.6**	**75.3**	**66.2**	**58.9**	**97.4**

et MinCq semblent eux aussi plus robustes). Nous observons que SVM, respectivement PV-SVM, implique de plus faibles performances que MinCq, respectivement PV-MinCq. Ces observations confirment l'intérêt de la prise en compte du désaccord entre votants. Ensuite, l'étiquetage basé sur la *perturbed variation* montre de meilleurs résultats que l'étiquetage basé sur les plus proches voisins (concernant le problème de la translation, le classifieur PPV étiquette tous les exemples cibles par la même classe). Contrairement à un étiquetage basé sur les PPV, le couplage induit par la *perturbed variation* permet de contrôler la divergence entre les domaines puisque il se focalise clairement dans les zones de densité plus élevée en retirant les points cibles non couplés, en d'autres termes dans les régions où les domaines sont "proches". Ces résultats confirment que la *perturbed variation* couplée à MinCq produit une solution intéressante à la tâche de l'adaptation de domaine.

5 Conclusion et discussion

Nous proposons PV-MinCq un algorithme d'adaptation de domaine pour apprendre un vote de majorité sur un ensemble de fonctions à valeurs réelles. PV-MinCq tire sa source de la théorie de l'adaptation de domaine PAC-Bayésienne et de MinCq, un algorithme d'apprentissage supervisé qui minimise la C-borne, majorante du risque du vote, qui permet le contrôle du désaccord entre les fonctions (un terme connu pour être crucial en adaptation de domaine). Dans un premier temps, nous proposons une nouvelle formulation de la C-borne qui permet la prise en considération d'une fonction d'étiquetage. Nous en déduisons l'algorithme PV-MinCq suivant.

(*i*) À l'aide des données sources étiquetées, nous étiquetons les données cibles dans les régions où les marginales source et cible sont proches. Cette procédure a l'originalité de tirer avantage de la *perturbed variation* entre les marginales source et cible, et à l'avantage d'être non itérative (contrairement à la plupart des méthodes d'auto-étiquetage).

(*ii*) En se fondant sur notre formulation de la C-borne, nous appliquons MinCq sur ces points auto-étiquetés.

Nous mettons ensuite en évidence la nécessité de contrôler le compromis entre faible *perturbed variation* et faible risque source, compromis que nous contrôlons lors de la validation des hyperparamètres. Finalement, les résultats empiriques sont prometteurs et suggèrent de nouvelles directions de recherches.

Afin de s'attaquer à des tâches réelles, une première direction intéressante concerne le design/l'apprentissage d'une $\epsilon(\mathcal{H})$-correcte distance (ou métrique) d pour définir un auto-étiquetage spécifique à la tâche. En effet, notre analyse de l'auto-étiquetage suggère le besoin d'une métrique d impliquant une mesure de similarité pertinente dans l'espace des votants.

De plus, PV-MinCq pourrait être utile dans le contexte de l'apprentissage multi-vues ou multi-modalités [10] [XTX13, Sun13]. En effet, dans une telle situation, une solution naturelle consiste à (*i*) apprendre un modèle pour chacune des différentes vues/modalités, (*ii*) apprendre [11] un vote de majorité sur l'ensemble de ces modèles. Ainsi, pour adapter un vote d'un corpus source vers un corpus cible, l'étape (*ii*) peut être réalisée par PV-MinCq, comme cela a été récemment fait pour MinCq [MHA14].

Enfin, nos résultats positifs posent la question de l'utilité de la *perturbed variation* pour identifier ou apprendre des attributs ou des points partagés par les domaines (comme dans [GGS13]).

Remerciements

This work was in parts funded by the European Research Council under the European Unions Seventh

10. Lorsqu'un document est représenté de différentes manières.
11. (*ii*) est parfois appelée *stacking* ou fusion de classifieurs.

Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

Références

[BBC+10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2) :151–175, 2010.

[BBCP07] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Proceedings of NIPS*, pages 137–144, 2007.

[BDU12] S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of ALT*, pages 139–153, 2012.

[BM10] L. Bruzzone and M. Marconcini. Domain adaptation problems : A DASVM classification technique and a circular validation strategy. *IEEE Transactions on PAMI*, 32(5) :770–787, 2010.

[Cat07] O. Catoni. *PAC-Bayesian supervised classification : the thermodynamics of statistical learning*, volume 56. Institute of Mathematical Statistic, 2007.

[GGS13] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks : Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of ICML*, 2013.

[GHLM13] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. PAC-Bayesian domain adaptation bound with specialization to linear classifiers. In *Proceedings of ICML*, 2013.

[GLLM09] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of ICML*, 2009.

[HM12] M. Harel and S. Mannor. The Perturbed Variation. In *Proceedings of NIPS*, pages 1943–1951, 2012.

[HPS13] A. Habrard, J.-P. Peyrache, and M. Sebban. Boosting for unsupervised domain adaptation. In *Proceedings of ECML-PKDD*, 2013.

[HSG+07] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. *Proceedings of NIPS*, 19 :601, 2007.

[Joa99] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of ICML*, pages 200–209, 1999.

[Lan05] J. Langford. Tutorial on practical prediction theory for classification. *JMLR*, 6 :273–306, 2005.

[LLM+07] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Proceedings of NIPS*, 2007.

[LMR11] F. Laviolette, M. Marchand, and J.-F. Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *Proceedings of ICML*, June 2011.

[Mar11] A. Margolis. A literature review of domain adaptation with unlabeled data. Technical report, University of Washington, 2011.

[McA99] D. A. McAllester. PAC-bayesian model averaging. In *Proceedings of COLT*, pages 164–170, 1999.

[McA03] D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of COLT*, pages 203–215, 2003.

[MHA12] E. Morvant, A. Habrard, and S. Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *KAIS*, 33(2) :309–349, 2012.

[MHA14] E. Morvant, A. Habrard, and S. Ayache. Majority Vote of Diverse Classifiers for Late Fusion. In *S+SSPR*, 2014.

[MMR08] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Proceedings of NIPS*, pages 1041–1048, 2008.

[PY10] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 22(10) :1345–1359, 2010.

[QCSSL09] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[See02] M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *JMLR*, 3 :233–269, 2002.

[Sun13] S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, pages 1–8, 2013.

[XTX13] Chang Xu, Dacheng Tao, and Chao Xu. A Survey on Multi-view Learning. Technical report, arXiv :1304.5634, April 2013.

[ZFY+10] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Proceedings of ECML-PKDD*, volume 6323 of *LNCS*, pages 547–562. Springer, 2010.