Étalonnage évidentiel actif de classifieurs SVM

Sébastien Ramel, Frédéric Pichon, François Delmotte

Étalonnage évidentiel actif de classifieurs SVM

Active evidential calibration of SVM classifiers

S. Ramel[1] F. Pichon[1] F. Delmotte [1]
[1] Univ. Artois, EA 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
F-62400 Béthune, France

{sebastien.ramel,frederic.pichon,francois.delmotte}@univ-artois.fr

Résumé :
Les méthodes évidentielles d'étalonnage de classifieurs binaires apportent une amélioration par rapport aux méthodes probabilistes, en représentant explicitement l'incertitude d'étalonnage induite par la quantité de données (étiquetées) d'apprentissage. Cette incertitude justifiée mais indésirable peut être réduite en ajoutant des données d'apprentissage, qui sont généralement coûteuses. D'où la nécessité de stratégies qui, étant donné un réservoir de données non étiquetées, sélectionneront des données intéressantes à étiqueter, c'est-à-dire celles induisant une baisse d'incertitude supérieure à la sélection aléatoire. Deux stratégies de ce type, inspirées de l'échantillonnage par incertitude en apprentissage actif, sont considérées dans cet article et appliquées à un ensemble de classifieurs SVM sur des jeux de données de classification binaire classiques. Les résultats expérimentaux montrent l'intérêt de l'approche vis-à-vis de la réduction de l'incertitude d'étalonnage, mais aussi vis-à-vis de l'amélioration des performances de classification.

Mots-clés :
Fonctions de croyance, étalonnage evidentiel, apprentissage actif.

Abstract:
Evidential calibration methods of binary classifiers improve upon probabilistic methods by representing explicitly the calibration uncertainty due to the amount of training (labelled) data. This justified yet undesirable uncertainty can be reduced by adding training data, which are in general costly. Hence the need for strategies that, given a pool of unlabelled data, will point to interesting data to be labelled, i.e., to data inducing a drop in uncertainty greater than a random selection. Two such strategies, inspired from uncertainty sampling in active learning, are considered in this paper and applied to an ensemble of SVM classifiers on some classical binary classification datasets. Experimental results show the interest of the approach to reduce calibration uncertainty, but also to improve classification performance.

Keywords:
Belief functions, evidential calibration, active learning.

1 Introduction

Les méthodes d'étalonnage probabilistes, telles que les régressions isotonique et logistique (méthode de Platt), permettent d'apprendre à partir de données d'apprentissage comment transformer les sorties d'un classifieur en probabilités qu'une instance appartienne à chacune des classes [1]. Elles sont utiles pour les nombreuses applications où il est important de fournir de telles probabilités plutôt que de simples décisions et où les classifieurs disponibles produisent des scores, comme les classifieurs SVM, ou des probabilités inexactes, comme les classifieurs Bayesiens naïfs [1, 2]. En outre, elles ont principalement été conçues jusqu'ici pour la classification binaire.

Une limitation de ces méthodes est qu'elles ne prennent pas en compte l'incertitude, due à la quantité de données d'apprentissage, dans leurs estimations des probabilités, et en particulier : moins il y a de données d'apprentissage, plus les probabilités estimées sont incertaines [3]. Afin de traiter cette difficulté, le problème de l'étalonnage a récemment été considéré dans le cadre de la théorie des fonctions de croyance, donnant lieu aux méthodes d'étalonnage dites évidentielles (voir [3] pour l'étalonnage d'un seul classifieur binaire et [4] pour l'étalonnage d'un ensemble de classifieurs binaires). Ces dernières méthodes sont capables de représenter explicitement l'incertitude due à la quantité de données d'apprentissage, ce qui est important

dans des domaines d'application critiques tels que la médecine et qui conduit également à de meilleures performances de classification que les méthodes d'étalonnage probabilistes comme montré dans [3, 4].

Bien qu'il soit important de représenter l'incertitude susmentionnée, il est encore plus intéressant que celle-ci soit aussi faible que possible. Afin de la réduire, il faut apporter des données (étiquetées) d'apprentissage supplémentaires, ce qui peut être coûteux et doit donc être effectué de manière efficace, c'est-à-dire tel que quel que soit le nombre de données étiquetées ajoutées, l'incertitude résultante soit réduite autant que possible. Il s'agit d'un problème similaire à celui de l'apprentissage actif [5], à la différence que l'objectif principal n'est pas d'améliorer les performances de classification mais plutôt de réduire l'incertitude, ce qui est le problème traité dans cet article.

Plus précisément, nous considérons la situation suivante : nous supposons un ensemble initial de données étiquetées à partir duquel des classifieurs peuvent être étalonnés de manière évidentielle, puis nous considérons qu'il est possible de demander itérativement à un oracle d'étiqueter des données d'un réservoir de données dont les étiquettes sont inconnues. Nous étudions deux stratégies pour décider quelles instances de ce réservoir doivent être soumises à l'oracle. Ces stratégies sont dans l'esprit de l'échantillonnage par incertitude (*uncertainty sampling*) de l'apprentissage actif [5], selon lequel les instances d'un réservoir sont ordonnées en fonction du degré d'incertitude qu'a le classifieur en ses prédictions quant à leur classe.

Cet article est organisé comme suit. Les rappels nécessaires concernant l'étalonnage evidentiel de classifieurs binaires sont fournis à la section 2. Ensuite, à la section 3, deux stratégies pour l'étalonnage évidentiel actif de tels classifieurs sont présentées et des résultats expérimentaux lorsque ces stratégies sont appliquées à des classifieurs binaires de type SVM sont rapportés. Enfin, des conclusions et perspectives sont données à la section 4.

2 Étalonnage évidentiel

L'étalonnage évidentiel de classifieurs binaires, introduit dans [3] dans le cas d'un classifieur unique et développé plus avant dans [4] pour un ensemble de classifieurs, repose sur des résultats récents de Kanjanatarakul *et al.* [6, 7] concernant la prédiction d'une variable aléatoire de Bernoulli. Ces résultats seront présentés après quelques rappels utiles sur la théorie des fonctions de croyance.

2.1 Rappels sur les fonctions de croyance

La théorie des fonctions de croyance permet de raisonner dans un contexte incertain. Considérons un cadre de discernement Ω, contenant toutes les réponses possibles à une question d'intérêt. Dans ce formalisme, l'incertitude vis-à-vis de la réponse à cette question est représentée par une fonction de masse m^Ω définie par l'application $m^\Omega : 2^\Omega \rightarrow [0,1]$ vérifiant $\sum_{A\subseteq\Omega} m^\Omega(A) = 1$. La quantité $m^\Omega(A)$ représente la part de croyance allouée seulement à l'hypothèse $A \subseteq \Omega$ et rien de plus spécifique. Tous les sous-ensembles $A \subseteq \Omega$ tels que $m^\Omega(A) > 0$ sont appelés élément focaux. Associées à la fonction m^Ω, les fonctions de croyance Bel^Ω et de plausibilité Pl^Ω définies par $Bel^\Omega(A) = \sum_{\emptyset\neq B\subseteq A} m^\Omega(B)$ et $Pl^\Omega(A) = \sum_{B\cap A\neq 0} m^\Omega(B)$, pour tout $A \subseteq \Omega$, représentent respectivement la part totale de croyance soutenant A, et la part de croyance ne contredisant pas A. Lorsque ses éléments focaux sont imbriqués, m^Ω est dite consonante et est caractérisée par sa fonction de contour pl^Ω définie par $pl^\Omega(\omega) = Pl^\Omega(\{\omega\}), \forall\omega \in \Omega$.

2.2 Prédiction d'une variable aléatoire de Bernouilli

Kanjanatarakul *et al.* [6, 7] ont proposé une approche générale qui, étant donné une connais-

sance sur un paramètre θ obtenue en observant une réalisation x d'une quantité aléatoire X de distribution $f_\theta(x)$ et représentée par une fonction de croyance Bel_x^Θ, permet de faire des déclarations sous la forme d'une fonction de croyance $Bel_x^\mathbb{Y}$ à propos d'une quantité aléatoire $Y \in \mathbb{Y}$ dont la distribution conditionnelle $g_{x,\theta}(y)$ étant donné $X = x$ dépend de θ. Dans cette approche, Bel_x^Θ doit être *induite par une source* [7] ; elle peut être obtenue par un certain nombre de méthodes évidentielles dédiées à l'inférence statistique, et en particulier la méthode évidentielle fondée sur la vraisemblance [8] auquel cas Bel_x^Θ est la fonction de croyance consonante dont la fonction de contour associée pl_x^Θ est la fonction de vraisemblance normalisée issue de l'observation x.

En particulier, si Y est une variable aléatoire binaire ($\mathbb{Y} = \{0, 1\}$) de distribution de Bernoulli $\mathcal{B}(\theta)$, $\theta \in [0, 1]$, et si Bel_x^Θ est une fonction de croyance consonante dont la fonction de contour associée pl_x^Θ est uni-modale et continue, nous avons [6] :

$$Bel_x^\mathbb{Y}(\{1\}) = \quad \hat{\theta} - \int_0^{\hat{\theta}} pl_x^\Theta(u)du, \qquad \text{(1a)}$$

$$Pl_x^\mathbb{Y}(\{1\}) = \quad \hat{\theta} + \int_{\hat{\theta}}^1 pl_x^\Theta(u)du, \qquad \text{(1b)}$$

où $\hat{\theta}$ maximise pl_x^Θ. Le degré de croyance $Bel_x^\mathbb{Y}(\{1\})$ représente la part de croyance supportant strictement $Y = 1$ alors que la plausibilité $Pl_x^\mathbb{Y}(\{1\}) = 1 - Bel_x^\mathbb{Y}(\{0\})$ est la part de croyance ne le contredisant pas. En outre, la différence $Pl_x^\mathbb{Y}(\{1\}) - Bel_x^\mathbb{Y}(\{1\})$, qui est égale à la masse $m_x^\mathbb{Y}(\{0, 1\})$ affectée à l'ignorance, est simplement l'aire sous la fonction pl_x^Θ et celle-ci tend vers 0 si, par exemple, X suit une distribution binomiale de paramètres n et θ, Bel_x^Θ est obtenue en utilisant la méthode basée sur la vraisemblance et n tend vers l'infini [6].

2.3 Méthodes d'étalonnage évidentielles

Soit $\mathcal{C} = \{(s_1, y_1), \ldots, (s_n, y_n)\}$ des données d'apprentissage dans le cadre d'un problème de classification binaire, où $s_i \in \mathbb{S}$ pour un domaine $\mathbb{S}$ est la sortie fournie par un classifieur pré-entrainé pour le i-ème exemple d'apprentissage d'étiquette $y_i \in \{0, 1\}$. Pour une instance de test de sortie $s \in \mathbb{S}$ et d'étiquette inconnue $y \in \{0, 1\}$, toute méthode évidentielle d'étalonnage proposée dans [3] renvoie deux valeurs : la croyance $Bel_{\mathcal{C},s}^\mathbb{Y}(\{1\})$ et la plausibilité $Pl_{\mathcal{C},s}^\mathbb{Y}(\{1\})$ que $y = 1$. Ces méthodes obtiennent ces deux valeurs en considérant l'étiquette y de l'instance de test comme la réalisation d'une variable aléatoire Y de distribution de Bernoulli $\mathcal{B}(\theta)$ étant donné la connaissance sur θ représentée par une fonction de croyance consonante $Bel_{\mathcal{C},s}^\Theta$ de fonction contour $pl_{\mathcal{C},s}^\Theta$ dépendant de $\mathcal{C}$ et s, et en appliquant alors à Y l'approche de prédiction rappelée dans la section 2.2.

La seule différence entre les méthodes d'étalonnage évidentielles présentées dans [3] est donc la façon dont $pl_{\mathcal{C},s}^\Theta$ est définie. Il y a en effet plusieurs façons de définir $pl_{\mathcal{C},s}^\Theta$: cela dépend de la méthode d'étalonnage probabiliste étendue et de l'approche évidentielle utilisée pour l'inférence statistique (voir [3, Section 4] pour plus de détails). Dans cet article, nous nous concentrons sur les méthodes d'étalonnage évidentielles où $pl_{\mathcal{C},s}^\Theta$ est obtenu en utilisant l'approche evidentielle basée sur la vraisemblance pour l'inférence statistique, puisque Xu *et al.* [3] ont montré que c'est l'approche présentant globalement les meilleures performances.

Plus précisément, considérons deux cas : $\mathbb{S} = \{0, 1\}$ et $\mathbb{S} = \mathbb{R}$. Le cas $\mathbb{S} = \{0, 1\}$ correspond à un classifieur renvoyant des sorties binaires et nous permettra d'étudier dans la section 3 les comportements de nos stratégies d'étalonnage évidentiel actif dans un cas simple. Le cas $\mathbb{S} = \mathbb{R}$ correspond à un classifieur renvoyant des scores, tel qu'un classifieur SVM, et nous permettra de rappeler progressivement la méthode d'étalonnage évidentielle sans doute la plus sophistiquée et performante considérée jusqu'à présent pour gérer un ensemble de classifieurs – les comportements de nos stratégies actives par rapport à cette dernière technique d'étalonnage

d'un ensemble de classifieurs seront également examinés dans la section 3.

Le cas $\mathbb{S} = \{0, 1\}$ peut être géré en utilisant l'extension évidentielle basée sur la vraisemblance de la méthode d'étalonnage probabiliste dite *binning* [3], auquel cas nous avons :

$$pl^{\Theta}_{\mathcal{C},s}(\theta) = \frac{\theta^{k_s}(1-\theta)^{n_s-k_s}}{\hat{\theta}_s^{k_s}(1-\hat{\theta}_s)^{n_s-k_s}}, \quad \forall s \in \mathbb{S}, \quad (2)$$

avec $k_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s, y_i = 1\}|$, $n_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s\}|$ et $\hat{\theta}_s = k_s/n_s$. L'équation (2) correspond à une approche binning dégénérée avec seulement deux bins et peut être dérivée rigoureusement sans se référer à l'étalonnage binning évidentiel, en suivant un raisonnement similaire à celui utilisé dans [3] pour obtenir ce dernier étalonnage.

Le cas $\mathbb{S} = \mathbb{R}$ peut être géré en utilisant l'extension évidentielle basée sur la vraisemblance de la régression logistique [3], auquel cas $pl^{\Theta}_{\mathcal{C},s}$ est défini comme suit, pour tout $s \in \mathbb{S}$:

$$pl^{\Theta}_{\mathcal{C},s}(\theta) = \sup_{\sigma_1 \in \mathbb{R}} pl^{\Sigma}_{\mathcal{C}}(\ln(\theta^{-1} - 1) - \sigma_1 s, \sigma_1), \quad (3)$$

avec $pl^{\Sigma}_{\mathcal{C}}(\sigma) = \frac{L(\sigma)}{L(\hat{\sigma})}$ pour tout $\sigma = (\sigma_0, \sigma_1) \in \Sigma = \mathbb{R}^2$, où $L(\sigma) = \prod_{i=1}^{n} p_i^{t_i}(1-p_i)^{1-t_i}$, avec $p_i = \frac{1}{1+\exp(\sigma_0+\sigma_1 s_i)}$ et $t_i = \frac{N_1+1}{N_1+2}$ si $y_i = 1$, $t_i = \frac{1}{N_0+2}$ si $y_i = 0$, avec $N_j = |\{(s_i, y_i) \in \mathcal{C} | y_i = j\}|$, et $\hat{\sigma}$ maximisant L.

Un intérêt particulier d'utiliser $pl^{\Theta}_{\mathcal{C},s}$ défini par (2) ou par (3) dans Eq. (1), est que $Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}) - Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}) = m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ diminue à mesure que n augmente [3]. En d'autres termes, $m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ reflète la quantité de données d'apprentissage, et en particulier moins il y a de données d'apprentissage, plus il y a d'ignorance ou d'incertitude.

Considérons maintenant un problème quelque peu plus complexe, où nous avons un ensemble de m classifieurs tels que, étant donnée une instance de test d'étiquette inconnue $y \in \{0, 1\}$, nous obtenons un vecteur de sorties $\mathbf{s} = (s^1, ..., s^m) \in \mathbb{R}^m$ avec s^j la sortie du j-ème classifieur. Pour pouvoir interpréter $\mathbf{s}$ par rapport à y, une solution proposée dans [4] consiste à étalonner *conjointement* les m classifieurs. L'étalonnage joint procède de façon similaire à l'étalonnage d'un seul classifieur : l'étiquette y est vue comme la réalisation d'une variable aléatoire avec une distribution de Bernoulli $\mathcal{B}(\theta)$ et une fonction de croyance $Bel^{\mathbb{Y}}_{\mathcal{C},\mathbf{s}}$ est dérivée en utilisant l'approche de prédiction (1) à partir d'une connaissance sur θ représentée par une fonction de contour $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$ dépendant de $\mathbf{s}$ et d'un ensemble d'apprentissage $\mathcal{C} = \{(\mathbf{s}_1, y_1), \ldots, (\mathbf{s}_n, y_n)\}$ où $\mathbf{s}_i$ est le vecteur de sorties fourni par les m classifieurs pour la i-ème instance d'apprentissage d'étiquette $y_i \in \{0, 1\}$. Plus spécifiquement, Minary *et al.* [4] ont proposé un étalonnage joint évidentiel correspondant à l'extension évidentielle basée sur la vraisemblance de la régression logistique multiple, qui est une généralisation de la régression logistique évidentielle rappelée ci-dessus et, en particulier, la définition de $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$ obtenue dans [4] est une simple généralisation multivariée de (3) (en raison d'un manque d'espace, nous renvoyons le lecteur à [4] pour la définition détaillée de $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$).

3 Étalonnage évidentiel actif

Comme nous l'avons vu, les méthodes d'étalonnage évidentielles retournent pour une instance de test de sortie s d'un classifieur, un degré de croyance $Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})$ et de plausibilité $Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})$ représentant, respectivement, la part de croyance soutenant strictement que l'étiquette y de l'instance est 1 et la part de croyance ne le contredisant pas. Par conséquent, plus grand est l'intervalle $[Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}), Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})]$, plus on est incertain quant au support qui doit être attribué à $y = 1$. Il est donc clair que, bien qu'il soit important que l'incertitude induite par les données d'apprentissage soit représentée, cette incertitude doit être suffisamment faible, sinon aucune conclusion utile quant à y ne peut être tirée, c'est-à-dire que le classifieur étalonné n'est pas utile.

Afin de réduire l'incertitude, il faut ajouter des données (étiquetées) d'apprentissage. Il est généralement possible et relativement aisé d'obtenir des données non étiquetées mais, selon le domaine, leur étiquetage peut être coûteux. En outre, il se peut que toutes les données d'apprentissage ne soient pas équivalentes par rapport à la baisse d'incertitude qu'elles induisent. Il semble donc utile de concevoir des stratégies qui, compte tenu d'un réservoir de données non étiquetées, indiqueront des données intéressantes à étiqueter, c'est-à-dire des données qui induiront une baisse d'incertitude supérieure à la sélection aléatoire dans le réservoir. De telles stratégies seront appelées par la suite stratégies d'étalonnage évidentiel actives, ou plus simplement stratégies actives, en opposition à la stratégie passive qu'est la sélection aléatoire. Nous proposons deux stratégies de ce type dans la section 3.1, que nous testons ensuite sur un seul classifieur puis sur un ensemble de classifieurs dans les sections 3.2 et 3.3.

3.1 Stratégies actives

Dans l'apprentissage actif basé sur un réservoir [5], un apprenant actif émet des requêtes sous la forme d'instances non étiquetées (provenant du réservoir) qui doivent être étiquetées par un oracle, puis ces instances nouvellement étiquetées sont déplacées vers l'ensemble d'apprentissage, avec l'objectif que la performance de classification soit améliorée plus rapidement qu'avec la stratégie de sélection aléatoire. Plusieurs stratégies de requête ont été proposées [5]. En particulier, l'échantillonnage par incertitude (*uncertainty sampling*) pour un classifieur avec des sorties probabilistes sélectionne l'instance non étiquetée du réservoir pour laquelle la sortie du classifieur a la plus grande entropie (de Shannon).

Puisque notre but est de réduire l'incertitude représentée par la quantité $m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ pour n'importe quelle instance de test de score $s \in \mathbb{S}$ donnée, une stratégie naturelle de requête consiste à sélectionner dans un réservoir $\mathcal{P} = \{s^{\mathcal{P}}_1, \ldots, s^{\mathcal{P}}_p\}$ d'instances non étiquetées avec sorties $s^{\mathcal{P}}_k$, $k = 1, \ldots, p$, du classifieur, l'instance $s^* \in \mathcal{P}$ qui a la plus grande incertitude $m^{\mathbb{Y}}_{\mathcal{C},s^*}(\{0,1\})$.

Nous notons qu'une mesure d'incertitude pour une fonction de masse $m^{\mathbb{Y}}$ est la mesure de Hartley généralisée [9], qui évalue sa non-spécificité et est définie comme suit : $GH(m^{\mathbb{Y}}) := \sum_{A \subseteq \mathbb{Y}} m^{\mathbb{Y}}(A) \log_2 |A|$; si $\mathbb{Y} = \{0,1\}$, nous avons $GH(m^{\mathbb{Y}}) = m^{\mathbb{Y}}(\{0,1\})$.

Cette stratégie naturelle est donc similaire à celle de l'échantillonnage par incertitude de l'apprentissage actif, sauf qu'elle utilise une autre mesure d'incertitude (l'entropie de Hartley généralisée au lieu de l'entropie de Shannon), et peut donc être appelée Échantillonnage de Hartley (EH) dans la suite. Cette stratégie sélectionne l'instance $s^*_{EH} \in \mathcal{P}$ telle que

$$s^*_{EH} = \arg\max_{s^{\mathcal{P}} \in \mathcal{P}} GH(m^{\mathbb{Y}}_{\mathcal{C},s^{\mathcal{P}}}) \qquad (4)$$

En plus de la stratégie EH, nous considérons à des fins de comparaison une autre stratégie de requête plus proche de l'échantillonnage par incertitude de l'apprentissage actif : cette seconde stratégie, appelée Échantillonnage Pignistique (EP), sélectionne l'instance $s^*_{EP} \in \mathcal{P}$ dont la distribution de probabilité pignistique associée [10] notée $BetP(m^{\mathbb{Y}}_{\mathcal{C},s^*_{EP}})$ a la plus grande entropie (de Shannon) :

$$s^*_{EP} = \arg\max_{s^{\mathcal{P}} \in \mathcal{P}} H(BetP(m^{\mathbb{Y}}_{\mathcal{C},s^{\mathcal{P}}})), \qquad (5)$$

avec $H(P)$ l'entropie de Shannon de la distribution de probabilité P. Notons que, puisque l'échantillonnage par incertitude est conçu pour améliorer les performances de classification, on peut s'attendre à ce que la stratégie EP améliore ce critère, mais il n'est pas clair que celle-ci améliore l'incertitude.

Par ailleurs, nous remarquons que la mesure de Hartley généralisée et l'entropie de Shannon de la transformation pignistique ont toutes deux

déjà montré leur intérêt pour améliorer les performances de classification dans le contexte de la classification active [11].

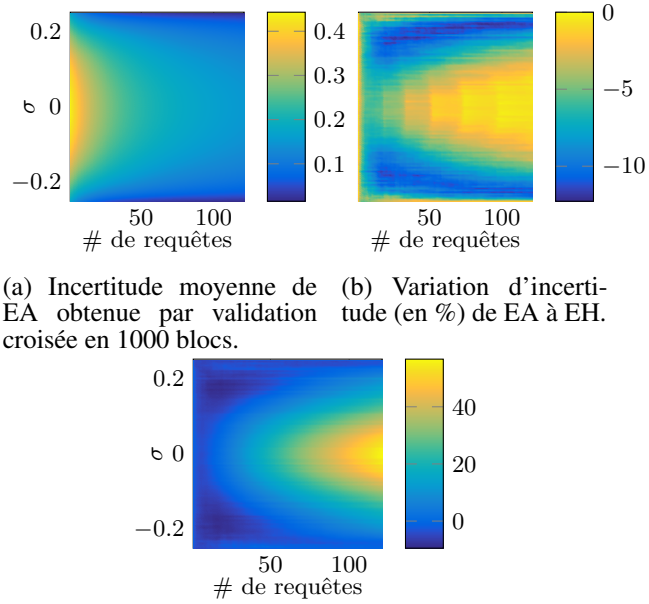
3.2 Étalonnage evidentiel actif d'un classifieur à sorties binaires

Les stratégies d'étalonnage actives décrites dans la section précédente sont d'abord testées sur un seul classifieur à sorties binaires, c'est-à-dire $\mathbb{S} = \{0, 1\}$, auquel cas le classifieur est étalonné en utilisant (2).

Le test est réalisé avec des données simulées. Plus précisément, soit $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, indiquant une distribution de Bernoulli bivariée donnée pour la paire (S, Y) de variables aléatoires binaires S et Y, où S représente la sortie du classifieur et Y la vraie classe. Une telle distribution est complètement caractérisée par les probabilités marginales $P(S = 1)$ et $P(Y = 1)$ et la covariance σ entre S et Y [12].

Dans notre expérience, nous avons choisi $P(S = 1) = P(Y = 1) = 0.5$, et considéré toutes les distributions possibles $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, ayant ces marginales : ce sont toutes les distributions qui sont obtenues en choisissant $\sigma \in [-0.25, 0.25]$, qui est la plage de valeurs possibles pour σ étant donné ces marginales.

Nous avons tiré 10^6 instances dans chacune de ces distributions jointes. Nous avons utilisé une procédure de validation croisée en 1000 blocs : les instances sont partitionnées aléatoirement en 1000 sous-ensembles. Chaque sous-ensemble (qui contient 1000 instances) est tour à tour considéré comme l'ensemble de test, alors que les autres sous-ensembles sont combinés pour obtenir un ensemble divisé aléatoirement en deux parties : la première composée de 10 instances est utilisée comme ensemble de données d'apprentissage initial $\mathcal{C}$ pour l'étalonnage évidentiel du classifieur, et la seconde composée des instances restantes est utilisée en qualité de réservoir. Le nombre



(a) Incertitude moyenne de EA obtenue par validation croisée en 1000 blocs.

(b) Variation d'incertitude (en %) de EA à EH.

(c) Variation d'incertitude (en %) de EA à EP.

Figure 1 – Comparaison des stratégies actives pour un classifieur à sortie binaire.

maximal de requêtes pour chaque stratégie de requête (EH, EP et Échantillonage Aléatoire (EA)) a été fixé à 120. Pour chaque sous-ensemble utilisé en qualité d'ensemble de test et pour chaque stratégie de requête, nous avons calculé la moyenne de l'incertitude, c'est à dire l'ignorance sur l'étiquette après l'étalonnage, des instances de test lorsque le nombre de requêtes augmente. Enfin, nous avons réalisé la moyenne de ces moyennes sur les 1000 sous-ensembles.

La figure 1 montre les performances en terme de réduction d'incertitude atteintes par les stratégies actives EH et EP par rapport à la stratégie passive (EA) utilisée comme référence. On peut noter que EH se comporte globalement mieux que EA (jusqu'à 12% mieux) - EH devient équivalent à EA quand σ se rapproche de 0 et le nombre de requêtes augmente, ou lorsque σ se rapproche de -0.25 et 0.25, représentant des situations de dépendance extrême entre S et Y. De plus, EP est bénéfique vis-à-vis de EA sur à peu près les mêmes zones que EH, quoique dans une moindre mesure. En revanche, EP est nettement nuisible (jusqu'à

55% pire) à mesure que le nombre de requêtes augmente et que σ se rapproche de 0. Notons que des figures similaires sont obtenues lorsque d'autres probabilités marginales $P(S = 1)$ et $P(Y = 1)$ sont utilisées (les figures sont alors des versions quelque peu déformées de celles présentées ici).

3.3 Étalonnage évidentiel actif de classifieurs SVM binaires

Les stratégies actives sont maintenant testées par rapport à un ensemble de 3 classifieurs SVM (entrainés avec la librairie LIBSVM) qui sont étalonnés conjointement en utilisant la régression logistique multiple évidentielle décrite dans la section 2.3.

Nous avons utilisé 6 jeux de données de classification binaires provenant du répertoire UCI : Australian, Heart, Ionosphere, Sonar, WDBC et Diabetes. Chaque jeu de données a été partitionné aléatoirement en 6 sous-ensembles : 3 sous-ensembles de 20 instances chacun pour entrainer chaque SVM, un sous-ensemble de 100 instances servant d'ensemble de test (excepté Sonar, pour lequel nous n'avons utilisé que 50 instances en raison de sa taille relativement petite), un sous-ensemble de 10 instances pour entraîner l'étalonnage joint initial des classifieurs et un sous-ensemble contenant les instances restantes agissant en qualité de réservoir.

Sur l'ensemble de test, nous avons calculé l'incertitude moyenne des stratégies EA, EH et EP, ainsi que leur score de Brier (erreur quadratique moyenne), qui est une mesure de performance de classification standard pour les méthodes d'étalonnage probabilistes [1, 2] (pour calculer ce score, nous avons transformé les fonctions de croyance produites par l'étalonnage évidentiel en distributions de probabilité en utilisant la transformation pignistique). Nous avons limité le nombre de requêtes à 20. L'ensemble du processus a été répété 100 fois avec partitionnement aléatoire, et les résultats obtenus ont été moyennés sur l'ensemble de ces répétitions.

Les moyennes d'incertitude sont présentées dans la figure 2 et les moyennes relatives au score de Brier sont affichées dans la figure 3. Comme dans l'expérience précédente, EH est meilleure que EP pour améliorer l'incertitude. En revanche, la stratégie EP est cette fois toujours meilleure que EA en terme d'incertitude. De plus, EH et EP sont meilleures ou équivalentes à EA en ce qui concerne le score de Brier, sauf sur le jeu de données Diabetes où EA est la meilleure stratégie. En moyenne, sur les six jeux de données considérés, EH est meilleure ou équivalente (pour tout nombre de requêtes entre 1 et 20) à EP, qui à son tour est meilleure que EA. Dans l'ensemble, cette expérience indique que les deux stratégies EH et EP peuvent améliorer l'incertitude ainsi que le score de Brier par rapport à EA, mais que EH peut être un meilleur choix que EP.

4 Conclusions

Dans cet article, les avantages de deux stratégies actives en matière de réduction de l'incertitude (et amélioration de la performance de classification) de l'étalonnage évidentiel de classifieurs binaires ont été étudiés. Des expériences préliminaires ont montré que bien que la stratégie d'échantillonnage pignistique puisse être bénéfique, celle-ci peut être surpassée par l'échantillonnage de Hartley. Des expériences plus approfondies (avec d'autres classifieurs, ensembles de données, méthodes d'étalonnage, tailles d'ensembles d'apprentissage et de réservoir) permettront d'affiner ces conclusions. Leur trouver des explications théoriques dans l'esprit de celles existantes en apprentissage actif [5] constitue une autre piste de travail intéressante. Enfin, nous envisageons d'appliquer cette approche à un système de détection de l'état d'un conducteur dont les données d'étalonnage sont coûteuses.

Remerciements :

Ce travail est financé en partie par le projet ELSAT2020, qui est cofinancé par l'Union européenne avec le Fonds européen de développement régional, l'État français et le Conseil régional des Hauts de France.

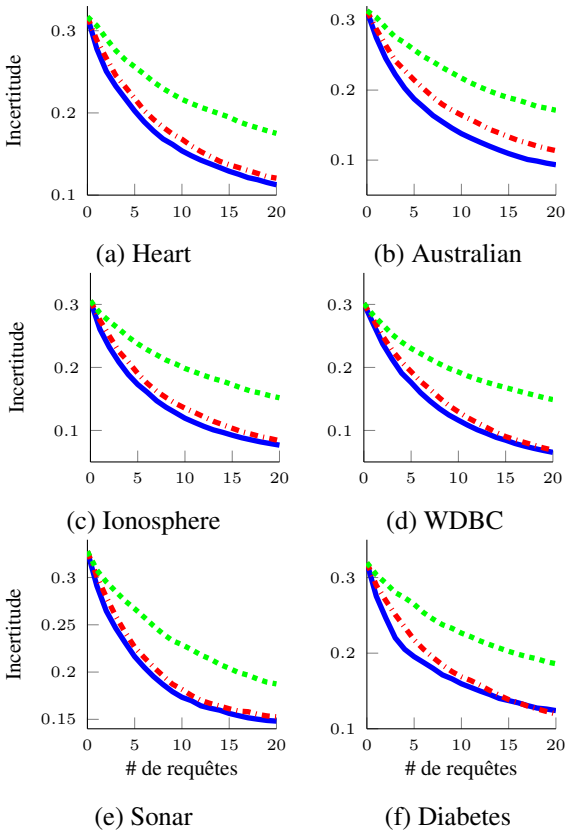


Figure 2 – Comparaison en terme d'incertitude des stratégies EH (bleu plein), EP (pointillés-point rouge) et EA (pointillés vert) pour un ensemble de 3 classifieurs SVM, sur 6 jeux de données UCI.

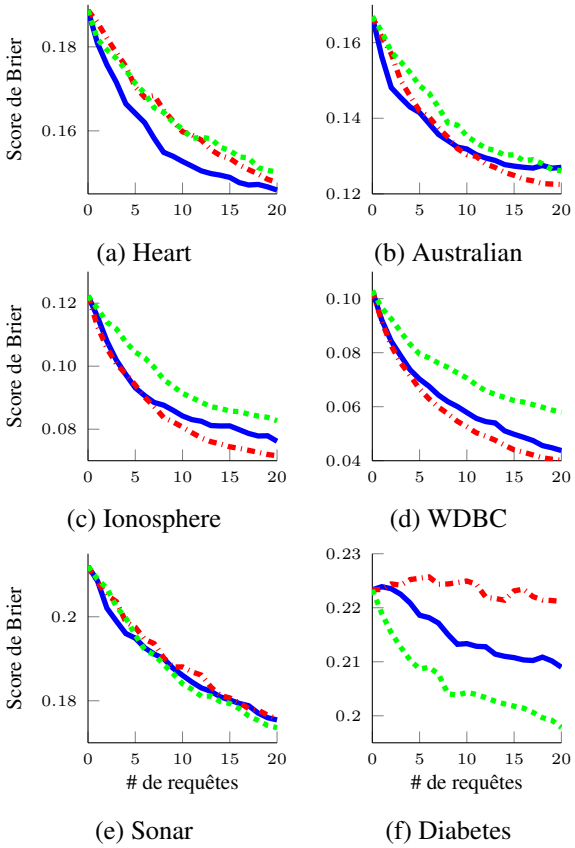


Figure 3 – Comparaison en terme de score de Brier des stratégies EH (bleu plein), EP (pointillés-point rouge) et EA (pointillés vert) pour un ensemble de 3 classifieurs SVM, sur 6 jeux de données UCI.

Références

[1] Niculescu-Mizil, A., Caruana, R. : Predicting good probabilities with supervised learning. Dans : Proc. of ICML. (2005) 625–632

[2] Zhong, W., Kwok, J.T. : Accurate probability calibration for multiple classifiers. Dans : Proc. of IJCAI. (2013) 1939–1945

[3] Xu, P., Davoine, F., Zha, H., Denoeux, T. : Evidential calibration of binary SVM classifiers. Int. J. Approx. Reason. **72** (2016) 55–70

[4] Minary, P., Pichon, F., Mercier, D., Lefevre, E., Droit, B. : Evidential joint calibration of binary SVM classifiers. Soft Computing. (à paraître)

[5] Settles, B. : Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)

[6] Kanjanatarakul, O., Sriboonchitta, S., Denœux, T. : Forecasting using belief functions : an application to marketing econometrics. Int. J. Approx. Reason. **55**(5) (2014) 1113–1128

[7] Kanjanatarakul, O., Denœux, T., Sriboonchitta, S. : Prediction of future observations using belief functions : A likelihood-based approach. Int. J. Approx. Reason. **72** (2016) 71–94

[8] Denoeux, T. : Likelihood-based belief function : Justification and some extensions to low-quality data. Int. J. Approx. Reason. **55**(7) (2014) 1535–1547

[9] Klir, G.J. : Uncertainty and Information : Foundations of Generalized Information Theory. John Wiley & Sons, Inc (2005)

[10] Smets, P., Kennes, R. : The transferable belief model. Artif. Intell. **66**(2) (1994) 191–234

[11] Reineking, T. : Active classification using belief functions and information gain maximization. Int. J. Approx. Reason. **72** (2016) 43–54

[12] Teugels, J.L. : Some representation of the multivariate Bernoulli and binomial distributions. J. Multivar. Anal. **32** (1990) 256–268