Transformation de scores SVM en fonctions de croyance

Philippe Xu[1] Franck Davoine[1,2] Thierry Denœux[1]

[1] UMR CNRS 7253, Heudiasyc, Université de Technologie de Compiègne, France
[2] LIAMA, CNRS, Key Lab of Machine Perception (MOE), Peking University, Chine

philippe.xu@hds.utc.fr

Résumé

La combinaison de plusieurs classifieurs, entrainés à partir de données ou caractéristiques distinctes, présente de nombreux intérêts pour les problèmes d'apprentissage supervisé. L'une des difficultés majeures de la combinaison est de représenter les sorties des différents classifieurs sous une forme commune, dans la plupart des cas sous la forme d'une probabilité à postériori. Cette transformation, appelée calibration, joue un rôle central dans la combinaison. Dans cet article, nous étendons les approches classiques de calibration probabilistes en utilisant la théorie des fonctions de croyance. Nous montrons, notamment, l'importance d'utiliser une borne inférieure et supérieure plutôt qu'une simple mesure probabiliste. Des résultats expérimentaux sur la transformation de scores SVM montrent l'apport des fonctions croyances.

Mots Clef

Combinaison de classifieurs, calibration de classifieurs, théorie de Dempster-Shafer, Support Vector Machines, fusion d'informations.

Abstract

The combination of many classifiers, trained using different data or features, is of great interest in machine learning. One main difficulty is to transform the outputs of the classifiers into a common representation, usually a class membership posterior probability. This transformation, called calibration, plays an important role in the combination. In this paper, we extend classical probabilistic calibration methods using the theory of belief functions. We show the importance of having a lower and upper bound instead of a single probability measure. Experimental results on the transformation of SVM scores show the gain of using belief functions.

Keywords

Classifier combination, classifier calibration, Dempster-Shafer theory, Support Vector Machines, information fusion.

1 Introduction

La combinaison de classifieurs est un problème important en apprentissage automatique. La combinaison de classifieurs, pouvant provenir de différents capteurs, données d'entrainement, modèles ou experts, donne souvent de meilleurs résultats que l'utilisation d'un unique classifieur. De manière générale, les méthodes de combinaison peuvent être séparées en deux types: les combinaisons entrainables et les combinaisons non-entrainables [1].

Dans le premier cas, les sorties des classifieurs à combiner sont données en entrée à un nouveau classifieur. Ce dernier peut alors être entrainé en utilisant des méthodes classiques d'apprentissage automatique. Les approches entrainables sont souvent préférées car elles peuvent être asymptotiquement optimales. Dans certains cas, les classifieurs de base sont spécifiquement construits pour être utilisés dans un tel schéma de combinaison, le *bagging* et le *boosting* en sont des exemples. L'un des principaux inconvénients de telles approches est qu'elles nécessitent un ré-entrainement. Cela demande non seulement d'avoir un jeu de données d'entrainement commun à tous les classifieurs mais cela limite également l'addition ultérieure de nouveaux classifieurs. En pratique, de nouvelles sources d'information peuvent arriver de manière itérative, de nouveaux capteurs peuvent être ajoutés ou de nouvelles données d'apprentissage peuvent devenir disponibles. Il devient alors contraignant de réapprendre un nouveau modèle à chaque fois.

Les approches non-entrainables consistent, elles, à combiner directement les sorties des classifieurs de base, en utilisant une règle de combinaison prédéfinie. L'exemple le plus simple est la combinaison par vote, c'est l'un des rares cas où il n'est pas nécessaire de traiter les sorties des classifieurs. En général, les sorties des différents classifieurs ne sont pas directement comparables, et à fortiori non combinables. Une étape de calibration, transformant les sorties des classifieurs en probabilités à postériori, devient alors nécessaire. Une calibration peut être également nécessaire même lorsque les sorties des classifieurs sont probabilistes, comme dans le cas du bayésien naïf, les probabilités d'appartenance estimées n'étant parfois pas assez précises [2]. La combinaison des probabilités peut alors se faire en utilisant des opérateurs simples comme le pro-

duit, la somme, le maximum, le minimum ou encore la médiane. Une méthode plus robuste consiste à utiliser une somme pondérée. Si la pondération est optimisée en utilisant tous les classifieurs, on retombe dans le cas des combinaisons entrainables. Dans le cas contraire, si chaque classifieur peut estimer son propre poids, alors on reste dans le cadre des combinaisons non-entrainables. L'utilisation d'approches non-entrainables peut conduire à des résultats sous-optimaux, les interactions, potentielles, entre les classifieurs n'étant pas considérées. En contrepartie, il devient très simple d'ajouter de nouveaux classifieurs.
Dans cet article, nous nous intéressons uniquement au second type de combinaison. Dans ce cas, les performances de chaque classifieur de base ne sont pas de la plus grande importance, par contre, l'étape de calibration devient cruciale [3, 4]. Nous centrons notre étude sur la calibration d'un classifieur SVM dans le cadre de problèmes de classification binaire. Dans un premier temps, nous passons en revue les méthodes existantes de calibration. Nous montrons, ensuite, les limites des approches probabilistes puis les étendons en utilisant la théorie des fonctions de croyance [5]. Enfin, nous validons notre approche par des résultats expérimentaux.

2 Calibration et combinaison

Étant donné des données d'entrainement $x_i \in \mathbb{R}^n$ de classe $y_i \in \{0, 1\}$, pour $i = 1, \ldots, k$, et un classifieur s, la calibration consiste à estimer la probabilité à postériori $P(y = 1|s(x))$, où $s(x) \in \mathbb{R}$ est le score donné par le classifieur s à un exemple de test x de classe y inconnue. De nombreuses méthodes de calibration peuvent être trouvées dans la littérature [6]. Parmi elles, les plus utilisées sont basées sur la régression logistique [7], le *binning* [8] et la régression isotonique [9].
L'approche de Platt [7] utilise une régression logistique sur les données d'entrainement. La transformation des scores SVM en probabilités se fait alors par une fonction sigmoïde. La méthode de *binning* [8] consiste à discrétiser l'espace des scores en plusieurs intervalles contigus. À chaque intervalle est alors associé une mesure de probabilité qui correspond au ratio d'exemples de classe positive sur le nombre total d'exemples dont les scores sont, effectivement, dans l'intervalle considéré. Enfin, la régression isotonique [9] utilise une fonction croissante constante par morceaux minimisant l'erreur quadratique moyenne.
La figure 1 illustre la calibration des scores d'un classifieur SVM entrainé sur le jeu de données *Australian* de la base de données UCI (Statlog) [1]. Pour la méthode de *binning*, les intervalles suivants ont été utilisés: $]-\infty, -3]$, $]-3, -2]$, $]-2, -1], \ldots,]+2, +3]$, $]+3, +\infty[$.
Ces approches probabilistes présentent plusieurs limitations. Tout d'abord, les incertitudes liées à l'étape de calibration, notamment celle liée au nombres d'exemples, n'est pas prise en compte. Par exemple, un intervalle comptant 10 exemples positifs sur 20 et un autre comptant 100

1. http://archive.ics.uci.edu/ml

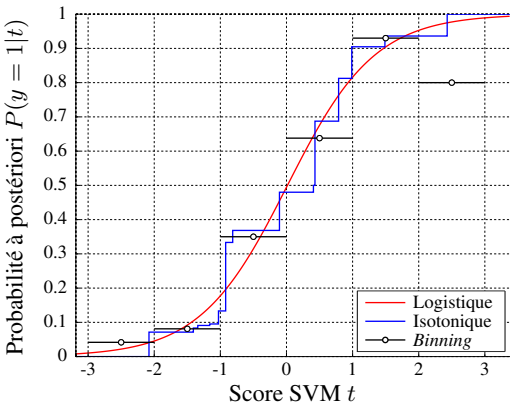


FIGURE 1 – Calibration de scores SVM sur le jeu de données *Australian*.

exemples positifs sur 200 seront tous les deux représentés par une probabilité de $1/2$. Or, dans le second cas de figure, l'estimation est beaucoup plus certaine. De même, les paramètres de la sigmoïde calculés par la régression logistique seront d'autant plus certains que le nombre de données d'entrainement sera important. Un autre point critiquable est le changement de la frontière de décision. En effet, un score SVM nul ne coïncide, en général, pas avec une probabilité de $1/2$ après calibration. Cela peut être vu comme un signe de sur-apprentissage.

Une fois les classifieurs calibrés, leurs sorties probabilistes peuvent être combinées. Les opérateurs de combinaison les plus communément utilisés sont le produit, la somme et la somme pondérée. Pour ce dernier, une pondération assez simple consiste à utiliser comme poids la précision propre à chaque classifieur. La précision peut être calculée par validation croisée à l'étape d'apprentissage. En théorie, si la calibration est parfaite, la pondération des classifieurs n'est pas nécessaire. En pratique, cette *double pondération* donne souvent de meilleurs résultats [6]. Cela montre, d'une certaine manière, que l'incertitude des classifieurs n'est pas totalement encodée par la calibration.

Des approches de normalisation de scores ne nécessitant pas une nouvelle phase d'apprentissage peuvent être trouver dans la littérature [10]. La normalisation obtenue est souvent assimilée à une mesure de probabilité mais ne sont souvent pas calibrés. Cela peut poser problème lorsque plusieurs classifieurs de différentes natures doivent être combinés.

3 Théorie des fonctions de croyance

La théorie de Dempster-Shafer [5], aussi appelée théorie des fonctions de croyance ou théorie de l'évidence, est une généralisation de la théorie des probabilités. Elle peut être utilisée à la fois pour la prédiction et l'inférence statistique.

3.1 Fonctions de croyance prédictives

Soit $\Omega = \{\omega_1, \ldots, \omega_K\}$ un ensemble de classes. Une *fonction de masse* est une fonction $m : 2^\Omega \to [0, 1]$ vérifiant:

$$\sum_{A \subseteq \Omega} m(A) = 1, \qquad m(\emptyset) = 0. \tag{1}$$

Étant donné un objet de classe $\omega \in \Omega$, la quantité $m(A)$, pour un ensemble $A \subseteq \Omega$, représente la croyance allouée spécifiquement à l'hypothèse $\omega \in A$. En particulier, la masse $m(\Omega)$ représente la quantité d'ignorance totale. Une fonction de masse peut être représentée de manière équivalente par une fonction de *croyance* ou de *plausibilité* définies, respectivement, par:

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \quad pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{2}$$

pour tout ensemble $A \subseteq \Omega$. La croyance $bel(A)$ représente la quantité d'évidence soutenant strictement l'hypothèse $\omega \in A$, tandis que la plausibilité $pl(A) = 1 - bel\left(\overline{A}\right)$ représente la quantité d'évidence ne la contredisant pas. Ces fonctions représentant une information sur la classe d'une observation seront dites prédictives.
Étant données deux fonctions de masse m_1 et m_2, issues de sources d'information indépendantes, elles peuvent être combinées en une nouvelle fonction de masse $m_{1,2}$ par la règle de combinaison de Dempster:

$$m_{1,2}(A) = \begin{cases} 0, \text{ si } A = \emptyset, \\ \frac{1}{1-\kappa} \sum\limits_{B \cap C = A} m_1(B) m_2(C), \text{ sinon,} \end{cases} \tag{3}$$

où $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ mesure le conflit entre les deux fonctions de masse.
La quantité d'incertitude intrinsèque d'une fonction de masse peut être accrue par l'utilisation d'un facteur d'affaiblissement $\delta \in [0, 1]$. Ce dernier peut être vu comme jouant un rôle équivalent aux poids dans une combinaison probabiliste pondérée. Il peut, cependant, être directement incorporé dans une fonction de masse affaiblie:

$${}^\delta m(A) = \begin{cases} (1-\delta) m(A), & \text{si } A \subsetneq \Omega, \\ (1-\delta) m(\Omega) + \delta, & \text{sinon.} \end{cases} \tag{4}$$

Si $\delta = 0$, la fonction de masse reste inchangée, tandis que si $\delta = 1$, la fonction de masse affaiblie sera l'ignorance totale ${}^\delta m(\Omega) = 1$. Cette dernière n'aura aucune influence lorsqu'elle sera combinée avec une autre fonction de masse.
Il est à noter qu'une distribution de probabilité est un type particulier de fonction de masse. Il est, toutefois, possible de la transformer en une fonction de masse non-probabiliste. La transformation pignistique inverse, proposée par Dubois *et al.* [11], se base sur le principe du minimum d'information et le q-ordonnancement. Dans le cadre d'un problème binaire, la fonction de croyance obtenue à partir d'une probabilité P est définie par:

$$bel_{PigInv}(\{1\}) = \begin{cases} 0 & \text{si } P(1) \leq \frac{1}{2}, \\ 2P(1) - 1 & \text{sinon,} \end{cases} \tag{5}$$

$$pl_{PigInv}(\{1\}) = \begin{cases} 2P(1) & \text{si } P(1) \leq \frac{1}{2}, \\ 1 & \text{sinon.} \end{cases} \tag{6}$$

La croyance et la plausibilité d'avoir un exemple positif engendrent, en fait, le plus grand intervalle centré en $P(1)$ possible.

3.2 Fonctions de croyance induites par inférence statistique

Les fonctions de croyance peuvent également être utilisées sur un domaine continu pour de l'inférence statistique. Denœux [12] a récemment justifié l'utilisation de la fonction de vraisemblance pour construire une fonction de croyance. Dans le cas d'un problème binaire, la probabilité qu'une observation x soit de classe positive $P(1|x)$, que nous appelerons également probabilité de succès, peut être écrite sous la forme d'une loi de Bernoulli de paramètre $\theta \in [0, 1]$, *i.e.* $P(1|x) = \theta$.
Étant donnée $L(\theta)$ une fonction de vraisemblance sur le paramètre θ, Denœux [12] propose d'utiliser la vraisemblance relative comme fonction de contour:

$$\forall \theta \in [0, 1],\ pl^\Theta(\theta) = \frac{L(\theta)}{\sup_{\theta' \in [0,1]} L(\theta')}, \tag{7}$$

à laquelle est associée la fonction de plausibilité suivante:

$$\forall A \subseteq [0, 1],\ Pl^\Theta(A) = \sup_{\theta \in A} pl^\Theta(\theta). \tag{8}$$

Lorsque la fonction de contour est uni-modale, ce qui sera souvent le cas en pratique, une fonction de croyance prédictive peut être engendrée par:

$$bel_{Vrai}(\{1\}) = \hat{\theta} - \int_0^{\hat{\theta}} pl^\Theta(u) du, \tag{9}$$

$$pl_{Vrai}(\{1\}) = \hat{\theta} + \int_{\hat{\theta}}^1 pl^\Theta(u) du, \tag{10}$$

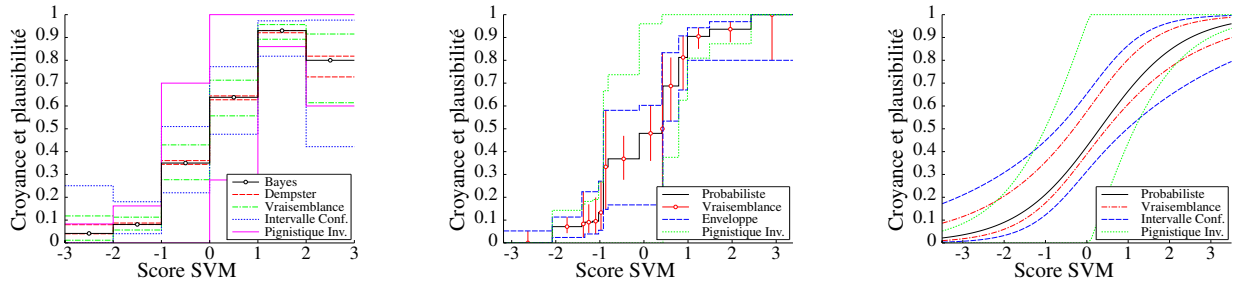
où $\hat{\theta}$ est la valeur de θ maximisant la fonction de vraisemblance. Ces quantités peuvent dans certains cas être calculées explicitement, sinon, des approches numériques doivent être utilisées.

4 Calibration évidentielle

Les méthodes de calibration probabilistes peuvent être étendues en utilisant la théorie des fonctions de croyance. Au lieu d'avoir une unique probabilité de succès $P(1)$, une croyance et une plausibilité forment respectivement une borne inférieure et supérieure de $P(1)$.

4.1 Binning

Dans le cas du *binning*, un intervalle contenant X exemples positifs sur N peut être vu comme le résultat d'une loi bi-



(a) *Binning* (b) Régression isotonique (c) Régression logistique

FIGURE 2 – Calibration évidentielle d'un classifieur SVM sur le jeu de données *Australian*.

nomiale. La probabilité de succès correspond alors au paramètre inconnu $\tau \in [0,1]$ de la loi de Bernoulli sous-jacente. D'un point de vue bayésien, la meilleure estimation est simplement:

$$P_B(1) = \hat{\tau} = \frac{X}{N}, \quad P_B(0) = 1 - \hat{\tau} = \frac{N-X}{N}. \tag{11}$$

Le modèle de Dempster, assez similaire à un estimateur de Laplace, donne la fonction de masse suivante:

$$m_D(\{1\}) = \frac{X}{N+1}, \quad m_D(\{0\}) = \frac{N-X}{N+1}. \tag{12}$$

Pour prendre en compte l'incertitude liée au nombre d'exemples, une approche communément considérée est l'utilisation d'intervalles de confiance. Un intervalle de confiance $[\underline{\tau}, \overline{\tau}]$ à un degré de confiance $1-\alpha \in [0,1]$, *i.e.* $P(\underline{\tau} \leq \tau \leq \overline{\tau}) = 1-\alpha$, peut être représenté par la fonction de contour suivante:

$$pl^T_{IntConf}(\tau) \quad = \quad \begin{cases} 1 & \text{if } \underline{\tau} \leq \tau \leq \overline{\tau}, \\ \alpha & \text{otherwise.} \end{cases} \tag{13}$$

Une autre manière d'obtenir une fonction de contour est l'utilisation de la fonction de vraisemblance, on obtient alors:

$$\forall \tau \in [0,1], \quad pl^T_{Vrai}(\tau) = \frac{\tau^X (1-\tau)^{N-X}}{\hat{\tau}^X (1-\hat{\tau})^{N-X}}. \tag{14}$$

Pour ces deux fonctions de contour (13-14), les fonctions définies par (9-10) peuvent être utilisées pour calculer une fonction de croyance prédictive.
La figure 2(a) montre les différentes fonctions de croyance obtenues en calibrant un classifieur SVM sur le jeu de données *Australian*. Pour l'approche basée sur les intervalles de confiance, la méthode exacte de Clopper-Pearson a été utilisée avec un degré de confiance de 95%. On remarque que le modèle de Dempster est très proche du modèle bayésien tandis que l'approche par intervalles de confiance est plus conservative. La méthode basée sur la vraisemblance est intermédiaire aux deux. La transformation pignistique inverse, quant à elle, ne prend pas en compte le nombre d'exemples, l'incertitude liée à certains intervalles de score, comme l'intervalle $]-1,0]$, paraît sur-estimée.

4.2 Régression isotonique

La calibration par régression isotonique donne, comme pour le *binning*, une discrétisation de l'espace des scores en intervalles. Les méthodes présentées précédemment peuvent alors également être utilisées. La figure 2(b) montre, pour chaque intervalle, les bornes inférieures et supérieures définies par une fonction de croyance basée sur la vraisemblance. Par contre, l'enveloppe induite par ces bornes n'est pas, en général, croissante. Il est toutefois très simple d'obtenir une enveloppe croissante, comme illustrée sur la figure 2(b).

4.3 Régression logistique

Pour la régression logistique, il y a deux paramètres inconnus $(\theta_0, \theta_1) \in \mathbb{R}^2$. Étant donné un score $t \in \mathbb{R}$, le paramètre d'intérêt $\gamma_t \in [0,1]$ est défini par:

$$\gamma_t = g_t(\theta_0, \theta_1) = \frac{1}{1+\exp(\theta_0 + \theta_1 t)}. \tag{15}$$

En formulant la régression logistique comme un modèle linéaire généralisé, les intervalles de confiance de Wald peuvent être utilisés pour calculer un intervalle de confiance $\left[\underline{\gamma_t}, \overline{\gamma_t}\right]$ sur γ_t. La fonction de contour (13) peut alors être utilisée.
Il est également possible de calculer une fonction de contour sur γ_t en passant par une fonction de croyance sur les paramètres θ_0 et θ_1. La fonction de vraisemblance sur les données d'entrainement est définie par:

$$L(\theta_0, \theta_1) \quad = \quad \prod_{i=1}^{k} p_i^{y_i} (1-p_i)^{1-y_i}, \tag{16}$$

avec $p_i = g_{s(x_i)}(\theta_0, \theta_1)$. La fonction de contour $pl^{\Theta_0 \times \Theta_1}$, donnée par (7), induit alors la fonction de plausibilité $Pl^{\Theta_0 \times \Theta_1}$ définie par (8). À partir de cette dernière, la fonction de contour pl^{Γ}_t sur la variable γ_t devient:

$$\forall \gamma \in [0,1], \quad pl^{\Gamma}_t(\gamma) = Pl^{\Theta_0 \times \Theta_1}(g_t^{-1}(\gamma)). \tag{17}$$

La fonction de contour $pl^{\Theta_0 \times \Theta_1}$ étant uni-modale, la quantité $pl^{\Gamma}_t(\gamma)$ peut être calculée par simple descente de gradient. La figure 2(c) illustre les différents résultats de calibrations évidentielles par régression logistique.

4.4 Affaiblissement et décision

Comme défini par (4), une fonction de masse peut être affaiblie par un facteur $\delta \in [0, 1]$. Ce facteur, jouant également de rôle de poids dans une somme pondérée, peut être assimilé à la précision du classifieur. Une estimation de la précision peut être obtenue par validation croisée.
Un autre aspect, déjà discuté dans la partie 2, est le changement de la frontière de décision après calibration. Dans le cas d'un classifieur SVM, seul le signe du score importe lors de la décision, la valeur du score ne pouvant représenter qu'une incertitude sur cette décision. Ainsi, lorsqu'un score est positif, il ne devrait y avoir aucune masse allouée au singleton $\{0\}$ car rien ne soutient explicitement l'hypothèse négative. De même, si le score est négatif, aucune masse ne devrait être allouée au singleton $\{1\}$.
Pour obtenir de telles fonctions de masse, il suffit en pratique, de transférer les masses incorrectement attribuées à l'ensemble d'ignorance $\{0, 1\}$. La croyance de succès sera alors nulle pour un score négatif, tandis que la plausibilité de succès sera de un avec un score positif.

5 Résultats expérimentaux

Une évaluation expérimentale a été menée sur plusieurs problèmes de classification binaire du dépôt UCI. Pour chaque jeu de données, trois classifieurs SVM ont été entrainés sur des jeux de données distinctes de différentes tailles. Les deux premiers ont été entrainés avec un nombre fixe de données, tandis que le troisième a été entrainé avec un nombre variable allant de dix à une centaine. La Table 1 donne les nombres de données utilisés pour l'apprentissage et le test. Pour chaque expérience, une validation croisée avec un échantillonnage de 5 a été utilisée pour obtenir à la fois les scores SVM et une estimation de la précision de chaque classifieur pour les combinaisons pondérées. Chaque expérience a été répétée 20 fois en tirant aléatoirement les données d'apprentissage et de test. La librairie LibSVM [2] a été utilisée pour apprendre les classifieurs.
Un test statistique a été conduit pour comparer les précisions moyennes obtenues par les 20 répétitions. Les données d'apprentissage et de test n'étant pas indépendantes d'une répétition à l'autre, un test de Student de comparaison de moyennes à partir d'échantillons appariés conduit souvent à un test trop *libéral*. Une version corrigée, plus *conservative*, est proposée dans [13].
Les résultats expérimentaux sont détaillés sur la Table 2. Les meilleurs performances sont dans la plupart des cas obtenues en utilisant une approche par fonctions de croyance en gardant les décisions inchangées.
Pour le *binning*, les résultats sont significativement meilleures en gardant les décisions inchangées pour le jeu de données *sonar* mais également pour *ionosphere* et *liver-disorders* lorsque le troisième classifieur est entrainé sur plus de données que les deux autres. Les résultats obtenus en utilisant le modèle de Dempster, un intervalle de confiance ou la vraisemblance ne sont pas significativement différents. Ils sont toutefois souvent meilleurs qu'une approche par transformation pignistique inverse. On observe également un gain de performance lorsqu'une pondération est utilisée, soit sous la forme d'une somme pondérée soit sous la forme d'un facteur d'affaiblissement.
Pour la régression isotonique et la régression logistique, aucune méthode ne donne de résultats significativement meilleurs avec le test statistique *conservatif*. On peut toutefois remarquer un certain gain en utilisant les approches par fonctions de croyance par rapport aux combinaisons probabilistes, notamment lorsque les décisions sont gardées inchangées. Contrairement au *binning*, l'utilisation d'une pondération pour la régression isotonique et la régression logistique ne semble pas toujours améliorer les résultats.
Lorsque les décisions sont maintenues inchangées, les trois méthodes de calibration donnent des résultats très similaires. Ce n'est pas le cas en utilisant une calibration probabiliste qui peut être souvent soumise à un fort sur-apprentissage.

	SVM #1	SVM #2	SVM #3	Test
autralian	30	70	10–200	390
diabetes	30	70	10–200	468
heart	20	40	10–140	70
ionosphare	20	40	10–190	101
liver-disorders	20	40	10–190	95
sonar	20	40	10–90	58

TABLE 1 – Nombres de données utilisées pour l'apprentissage et le test.

6 Conclusion

Dans cet article, nous avons montré comment étendre les méthodes de calibration probabilistes en utilisant des fonctions de croyance. Ces dernières permettent de mieux représenter l'incertitude liée à la calibration. De plus, il est possible de garder les frontières de décision des classifieurs de base inchangées, ce qui évite le sur-apprentissage à l'étape de calibration.
Les méthodes de calibration proposées peuvent également être utilisées pour calibrer d'autres types de classifieurs. Ces approches peuvent aussi être étendues à des problèmes multi-classes en utilisant une décomposition en problèmes binaires du type un-contre-un ou un-contre-tous. Une comparaison entre les approches probabilistes et éventielles dans le cadre multi-classes sera l'objet de travaux futurs.
Également, ces méthodes nous serviront pour nos travaux en cours portant sur la fusion multi-capteurs pour la compréhension de scènes de conduite automobile.

Remerciements

Ce travail mené dans le cadre du Labex MS2T s'insère dans le programme Investissements d'Avenir géré par l'Agence Nationale de la Recherche (ANR-11-IDEX-0004-02). Il est également financé par le projet franco-chinois *Blanc International* ANR-NSFC PRETIV (ANR-11-IS03-0001).

2. http://www.csie.ntu.edu.tw/~cjlin/libsvm

		Jeu de données	australian			diabetes			heart			ionosphere			liver-disorders			sonar		
		Nb. de données	10	50	200	10	50	200	10	50	140	10	50	190	10	50	190	10	50	90
Binning	Prob.	Produit	78.6	84.2	83.4	57.5	69.1	69.7	75.0	79.3	76.9	59.7	75.2	81.8	46.4	50.1	55.3	61.1	64.2	72.9
		Somme	82.5	84.8	84.2	65.7	70.3	70.7	75.2	80.1	78.3	77.1	83.3	90.5	53.4	52.9	58.3	60.2	64.1	75.3
		Somme pondérée	83.1	84.7	84.3	67.2	70.7	70.6	76.6	80.5	81.0	82.8	85.1	91.6	54.6	55.7	59.4	62.2	66.7	78.3
	Fonctions de croyance	Pignistique Inv.	78.6	84.2	83.3	57.5	69.3	69.8	75.0	79.3	76.9	59.7	75.2	81.8	46.4	50.1	55.4	61.2	64.2	73.0
		Pignistique Inv.*	83.5	84.5	84.3	67.3	70.7	70.7	77.0	80.4	80.6	83.5	85.0	93.0	54.9	55.9	59.7	63.8	67.5	79.1
		Dempster	84.0	84.7	83.2	67.6	70.6	70.7	74.4	77.7	79.8	81.0	84.2	93.4	52.1	53.3	60.1	62.9	65.6	78.0
		Dempster*	84.2	84.8	84.2	68.5	70.9	71.2	78.0	80.9	81.9	84.8	85.3	93.8	54.4	55.6	62.7	63.4	67.5	80.3
		Intervalle Conf.	84.4	85.0	81.5	67.7	70.6	71.2	67.0	76.0	81.1	83.7	84.2	93.6	52.3	54.4	65.4	62.4	66.4	79.0
		Intervalle Conf.*	84.4	84.6	82.8	67.9	70.7	71.5	75.4	80.4	82.1	84.8	84.1	93.6	52.2	55.3	**66.5**	63.3	67.5	80.1
		Vraisemblance	83.7	84.6	82.9	67.9	70.7	71.2	73.9	77.6	79.1	82.0	84.3	93.3	54.4	53.9	59.2	62.8	63.9	77.1
		Vraisemblance*	84.2	84.7	84.2	68.3	70.9	71.1	77.6	80.7	81.4	84.6	85.0	93.6	54.4	55.3	61.8	63.3	67.2	80.0
	Décision inchangée	Pignistique Inv.	82.2	84.6	84.9	64.8	70.4	70.6	80.1	**81.7**	81.0	71.8	79.6	85.8	48.3	52.6	59.5	67.5	70.5	77.5
		Pignistique Inv.*	84.6	84.8	**85.1**	68.7	71.6	71.3	79.9	**81.2**	81.7	85.7	85.3	92.8	53.7	54.6	61.4	68.6	70.9	80.3
		Dempster	84.6	**85.3**	**85.2**	69.5	71.6	72.1	**80.6**	**81.3**	**82.3**	84.2	**85.6**	94.0	56.2	**57.5**	64.4	69.0	**72.4**	80.3
		Dempster*	**84.9**	**85.1**	**85.1**	**70.6**	**72.2**	**72.4**	**81.4**	**81.0**	**82.8**	**87.1**	**86.6**	**95.1**	**58.2**	**58.2**	65.4	**71.3**	**73.0**	**82.3**
		Intervalle Conf.	84.7	**85.2**	**85.1**	**70.2**	**71.8**	**72.7**	**80.6**	**81.4**	**82.6**	**86.0**	**85.8**	**94.9**	56.1	56.4	**66.6**	69.0	**72.2**	**82.8**
		Intervalle Conf.*	**84.8**	85.0	**85.2**	**70.4**	**72.2**	**72.6**	**81.1**	**81.7**	**82.7**	**86.7**	**85.9**	**95.3**	56.5	**57.1**	**67.3**	**70.7**	**72.6**	**82.8**
		Vraisemblance	**84.8**	**85.2**	**85.2**	69.2	**71.9**	72.1	80.1	80.9	**82.4**	83.8	**86.0**	93.8	56.7	**57.6**	62.9	68.0	71.9	80.3
		Vraisemblance*	**84.7**	**85.2**	**85.1**	69.9	**72.0**	72.1	79.8	**81.4**	**82.4**	**87.0**	**86.3**	94.5	**57.4**	**57.8**	64.2	**70.2**	**72.5**	82.1
Régression isotonique	Prob.	Produit	84.0	84.6	84.9	66.7	70.8	71.3	80.0	**80.7**	**82.4**	86.1	**86.5**	91.1	56.6	56.8	60.8	67.6	73.3	79.8
		Somme	84.0	84.7	84.9	67.3	71.0	71.7	79.6	**80.9**	**82.5**	85.6	84.7	91.5	57.2	57.8	61.5	67.8	73.4	80.0
		Somme pondérée	84.0	84.5	84.9	67.4	70.9	71.6	79.3	**80.4**	**82.2**	86.0	84.7	91.5	57.2	**59.3**	61.8	69.1	73.2	80.8
	Fonctions de croyance	Pignistique Inv.	83.9	84.5	84.9	66.7	70.7	71.3	80.0	**80.6**	**82.3**	86.1	**86.5**	91.7	55.9	56.8	60.5	67.6	73.1	79.7
		Pignistique Inv.*	83.7	84.5	84.9	67.4	71.0	71.4	80.0	**80.3**	82.1	85.8	85.6	93.3	57.1	**59.3**	61.4	68.7	72.8	80.9
		Dempster	84.1	84.7	85.0	68.2	70.9	71.6	80.0	**80.5**	81.6	85.5	85.5	92.3	58.1	**58.6**	63.6	67.1	72.8	80.1
		Dempster*	83.7	84.5	84.9	68.2	70.6	71.4	79.9	**80.5**	**82.5**	85.8	84.7	93.3	57.8	**59.6**	63.3	68.4	72.3	81.5
		Intervalle Conf.	83.9	84.5	84.9	69.0	70.7	70.8	77.9	79.8	**82.3**	86.0	84.5	92.7	**60.1**	58.0	**68.0**	67.0	69.8	80.6
		Intervalle Conf.*	84.0	84.5	85.0	68.4	70.3	71.2	80.1	**80.6**	**82.6**	85.2	84.0	93.2	**60.0**	**59.7**	**67.2**	69.0	72.2	81.2
		Vraisemblance	83.9	84.6	85.0	68.4	70.8	**71.8**	79.9	**80.2**	81.4	**87.4**	**86.2**	92.8	58.1	**58.9**	64.7	66.6	71.1	79.8
		Vraisemblance*	83.8	84.6	85.0	68.2	70.9	71.6	80.1	**80.4**	**82.4**	86.2	85.1	93.1	58.1	**59.6**	63.5	68.1	72.4	81.5
	Décision inchangée	Pignistique Inv.	84.2	84.7	85.1	67.6	71.4	71.5	**80.4**	**80.4**	82.1	**86.9**	**87.3**	92.3	55.6	56.7	61.1	**72.2**	73.7	80.6
		Pignistique Inv.*	84.1	84.7	85.1	68.2	**71.6**	**71.8**	79.9	**80.2**	81.9	**87.1**	**86.7**	**93.9**	56.8	58.3	61.7	**72.0**	73.4	**81.7**
		Dempster	84.6	**85.2**	**85.2**	**69.8**	**71.9**	**72.2**	79.8	**80.2**	**82.8**	86.4	**86.2**	93.2	**59.9**	**59.2**	64.5	71.0	73.0	80.3
		Dempster*	84.2	84.9	**85.2**	**69.7**	**72.1**	**72.2**	80.0	**80.1**	**82.6**	**87.4**	**86.2**	**94.3**	**59.4**	**59.1**	65.8	**72.8**	**74.1**	**82.2**
		Intervalle Conf.	**85.1**	84.9	**85.3**	**69.7**	71.3	**71.8**	**81.1**	**80.4**	**82.5**	86.4	85.2	93.1	**60.1**	**59.5**	**68.2**	71.5	73.4	80.4
		Intervalle Conf.*	84.4	**85.0**	**85.3**	**69.9**	71.2	71.7	**81.0**	**80.2**	**82.6**	86.1	84.5	**93.9**	**58.9**	**58.6**	**67.3**	**72.1**	73.4	**82.2**
		Vraisemblance	84.4	84.9	**85.3**	**69.9**	**71.8**	**72.2**	80.1	**80.1**	**82.4**	**87.3**	**86.0**	93.0	**59.3**	**58.7**	65.2	71.6	73.7	80.5
		Vraisemblance*	84.2	84.9	**85.2**	**69.4**	**71.8**	**71.9**	79.9	80.0	**82.5**	**87.7**	**86.4**	**93.9**	**59.5**	**58.6**	65.6	**72.8**	**74.9**	**81.8**
Régression logistique	Prob.	Produit	77.6	82.3	84.8	68.8	72.5	72.9	77.1	79.1	**82.7**	**88.6**	**87.9**	94.7	55.4	**58.2**	62.8	71.5	73.4	81.8
		Somme	77.7	82.3	84.9	68.8	72.5	72.9	77.1	79.0	**82.6**	**88.6**	**87.8**	94.5	55.3	**58.1**	62.6	71.5	73.4	81.7
		Somme pondérée	78.2	83.7	84.9	68.9	72.5	72.8	75.9	78.7	**82.6**	**88.2**	**87.5**	94.6	55.5	**57.5**	63.1	70.9	**74.0**	**82.6**
	Fc. de croyance	Pignistique Inv.	77.4	82.4	84.8	68.7	72.6	**73.0**	77.2	79.4	**82.8**	**88.6**	**87.9**	94.7	55.3	**58.0**	62.6	71.5	73.3	81.8
		Pignistique Inv.*	78.9	83.8	84.9	69.1	72.4	72.9	76.1	78.7	**82.6**	**88.0**	**87.4**	94.8	56.0	**57.5**	63.3	71.0	**73.8**	**82.3**
		Intervalle Conf.	78.1	82.9	84.9	69.6	**72.6**	72.9	77.5	79.3	**82.7**	87.4	**86.9**	94.8	55.0	57.1	64.7	71.6	73.7	**82.9**
		Intervalle Conf.*	80.2	83.9	84.9	69.9	72.4	**73.0**	76.2	78.5	**82.6**	87.3	**86.6**	95.0	56.3	56.6	64.1	70.9	73.4	**82.7**
		Vraisemblance	77.9	82.7	84.9	69.4	**72.7**	**73.0**	77.4	79.1	**82.7**	**87.8**	**87.3**	94.8	55.3	**58.2**	63.7	71.4	**73.9**	**82.7**
		Vraisemblance*	80.0	83.9	84.7	69.7	72.5	**72.9**	76.3	78.7	**82.5**	87.5	**87.0**	95.0	56.2	57.1	63.5	70.7	73.5	**82.7**
	Dec. inchangée	Pignistique Inv.	83.3	84.1	84.9	70.5	**72.7**	**73.3**	78.4	80.5	**82.9**	**88.9**	**87.9**	94.8	56.1	**57.4**	63.1	**72.6**	**74.7**	82.0
		Pignistique Inv.*	83.3	84.1	84.9	70.4	**72.6**	**73.0**	78.1	80.6	**82.8**	**88.4**	**87.4**	95.0	56.3	**57.7**	63.5	**72.4**	**75.0**	**82.2**
		Intervalle Conf.	**84.6**	**85.2**	**85.2**	**71.4**	72.5	72.7	**81.1**	**81.6**	**83.1**	86.0	**87.3**	95.0	**57.9**	56.9	64.6	71.6	72.5	81.9
		Intervalle Conf.*	84.1	85.0	**85.1**	**71.5**	**73.1**	**73.2**	**81.1**	80.9	**82.8**	**88.0**	**86.9**	**95.4**	56.8	57.0	65.0	**73.4**	**75.0**	**82.7**
		Vraisemblance	**84.7**	**85.3**	**85.2**	**71.3**	72.5	72.7	**81.6**	**81.9**	**82.9**	87.7	**87.4**	94.9	**59.0**	**58.8**	65.0	71.6	72.9	81.6
		Vraisemblance*	84.3	**85.1**	**85.2**	**71.3**	**72.9**	**73.2**	**81.0**	**81.2**	**82.6**	**88.5**	**87.5**	**95.2**	**58.7**	**58.9**	**66.1**	**72.8**	**74.7**	**82.5**

TABLE 2 – Précisions moyennes après calibration et combinaison des classifieurs SVM. La deuxième ligne (Nb. de données) correspond au nombre d'exemples utilisés pour entrainer le troisième classifieur. Les nombres en gras soulignés représentent les meilleures performances. Ceux uniquement en gras représentent les résultats n'étant pas significativement différents du meilleur par un test de Student avec un seuil de 5%. Ceux uniquement soulignés correspondent à la version corrigée. Les méthodes marquées du symbole $*$ sont celles utilisant un facteur d'affaiblissement.

Références

[1] R. P. Duin, "The combining classifier: to train or not to train?," in *ICPR*, vol. 2, pp. 765–770, 2002.

[2] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," in *ICML*, pp. 105–112, 1996.

[3] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *The Statistician*, vol. 32, no. 1, pp. 12–22, 1982.

[4] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, "On the effect of calibration in classifier combination," *Applied Intelligence*, vol. 38, no. 4, pp. 566–585, 2013.

[5] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[6] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, "Calibration of machine learning models," in *Handbook of Research on Machine Learning Applications and Trends*, pp. 128–146, IGI Global, 2009.

[7] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large-Margin Classifiers*, pp. 61–74, MIT Press, 1999.

[8] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *ICML*, pp. 609–616, 2001.

[9] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

[10] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *CVPR*, pp. 2933–2940, 2012.

[11] D. Dubois, H. Prade, and P. Smets, "A definition of subjective possibility," *International Journal of Approximate Reasoning*, vol. 48, pp. 352–364, 2008.

[12] T. Denœux, "Likelihood-based belief function: justification and some extensions to low-quality data," *International Journal of Approximate Reasoning (in press)*, 2014.

[13] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, pp. 239–281, 2003.