Vers un désenchevêtrement de l'ambiguïté de la tâche et de l'incertitude du modèle pour la classification avec option de rejet à l'aide de réseaux neuronaux

Titouan Lorieul, Alexis Joly

Vers un désenchevêtrement de l'ambiguïté de la tâche et de l'incertitude du modèle pour la classification avec option de rejet à l'aide de réseaux neuronaux

T. Lorieul[1] A. Joly[1]

[1] Zenith, LIRMM, Université de Montpellier, Inria, France

titouan.lorieul@gmail.com

Résumé

La classification avec option de rejet est un moyen d'aborder le problème de l'estimation de l'incertitude d'un classifieur. Les approches récentes s'attaquant à ce problème utilisent des critères basés sur une mesure, soit, de confiance, soit, de dispersion. Cependant, aucune d'entre elles ne combine explicitement les deux principales sources d'incertitude : l'ambiguïté de la tâche, intrinsèque à celle-ci, et l'incertitude du modèle, découlant de l'échantillonnage des données et de la stochasticité de l'apprentissage. Dans cet article, nous explorons comment ces deux quantités peuvent être fusionnées afin d'établir des critères de rejet plus efficaces. En particulier, nous proposons une série de méthodes combinant des mesures de désaccord et des estimations de l'ambiguïté en utilisant un ensemble de modèles. Des expériences sur des jeux de données synthétiques construits pour modéliser différents types d'incertitudes indiquent que ces nouveaux critères ont des performances similaires aux méthodes de référence. Néanmoins, des analyses plus approfondies montrent des indices empiriques qui mettent en avant l'existence d'information supplémentaire dans la distribution des résultats de l'ensemble. Dans les faits, le réjecteur idéal peut être une fonction plus complexe que les critères précédents, et peut même parfois être contre-intuitif.

Mots Clef

Classification avec option de rejet, estimation d'incertitude, réseaux neuronaux, ensemble de modèles.

Abstract

Classification with reject option is a way to address the problem of estimating the uncertainty of a classifier. Recent approaches to this problem use criteria based on either a confidence or a dispersion measure. However, they do not explicitly combine the two main sources of uncertainty : the ambiguity of the task, inherent to it, and the uncertainty of the model, resulting from data sampling and stochasticity of learning process. In this article, we explore how these two quantities can be merged to build more effective rejection criteria. In particular, we propose methods for combining disagreement measures and ambiguity estimates using an ensemble of models. Experiments on synthetic data sets constructed to model different types of uncertainties indicate that these new criteria have similar performance to the baselines. Nevertheless, more in-depth analyses show empirical evidence that highlights the existence of additional information in the distribution of the overall results. In practice, the ideal rejector may be a more complex function than the previous criteria, and may even be counter-intuitive at times.

Keywords

Classification with reject option, uncertainty estimation, neural networks, ensembles.

1 Introduction

Il est important de disposer de mesures précises de l'incertitude des prévisions d'un modèle dans de nombreux scénarios pratiques où l'on ne peut pas se permettre de commettre des erreurs. Cela est en particulier vrai dans des applications médicales, de conduite autonome, etc. Cependant, quantifier précisément cette information d'incertitude est un problème difficile, surtout lorsque le processus d'apprentissage n'est pas entièrement compris comme cela est le cas pour les réseaux neuronaux. Une façon d'assouplir cet objectif ambitieux, tout en progressant dans cette direction, est de permettre aux classifieurs de refuser de donner une réponse pour une entrée donnée. Ceci est connu sous le nom de *classification avec une option de rejet* [5, 14]. Cette décision de rejeter peut tirer parti de l'incertitude prédictive sans avoir à la modéliser complètement et ainsi permettre de mieux comprendre ce qui est nécessaire pour construire des modèles fournissant des informations d'incertitude plus complètes.

La classification avec option de rejet consiste à fournir deux fonctions : un prédicteur $h : \mathcal{X} \to \mathcal{Y}$ et un réjecteur $r : \mathcal{X} \to \{0, 1\}$, où $\mathcal{X}$ et $\mathcal{Y}$ sont, respectivement, les espaces d'entrée et de sortie de la tâche. La fonction apprise est alors

$$(h, r)(x) = \begin{cases} h(x) & \text{si } r(x) = 0, \\ \text{®} & \text{sinon (r(x) = 1),} \end{cases}$$

où ® désigne le refus de répondre.

Parmi les approches classiques du rejet, on peut distinguer deux grandes familles de méthodes. La première considère que h ne donne pas seulement une prédiction en $\mathcal{Y}$ mais fournit également une forme d'information de confiance, par exemple dans $\mathbb{R}$, qui peut être utilisée pour effectuer un rejet en seuillant sur sa valeur [1]. C'est le cas, par exemple, de la régression logistique et des réseaux de neurones qui donnent une distribution de probabilité catégorielle en sortie, ou des SVMs qui fournissent une distance à la frontière de décision. Nous les qualifierons de méthodes *basées sur la confiance*. L'autre catégorie tente de mesurer les fluctuations du classifieur et favorise le rejet dans les zones de l'espace d'entrée où sa variabilité est la plus élevée. L'hypothèse avancée par ces méthodes est que le prédicteur est plus susceptible d'être incertain dans sa décision dans ces zones-là. Ces approches ont été particulièrement explorées dans le contexte de l'apprentissage actif [21] et dans les réseaux neuronaux bayésiens [9]. Nous les qualifierons de méthodes *basées sur la dispersion*.

Nous affirmons que, en fait, ces deux approches sont incomplètes dans le sens où elles ne saisissent que partiellement l'information sur l'incertitude. En effet, cette dernière peut être divisée en deux catégories : *l'ambiguïté de la tâche* et *l'incertitude du modèle*. La première est intrinsèque à la tâche que nous voulons accomplir. Par exemple, une image de la fleur d'une plante ne contient qu'une information partielle qui peut ne pas être suffisante pour distinguer deux espèces similaires. Cette incertitude provient du bruit de la mesure (prendre une photo dans notre exemple) ainsi que du bruit dans le processus d'annotation. Elle est, en fait, directement liée à la fonction de régression

$$\eta_k(x) = \Pr\left[Y = k \mid X = x\right]$$

de l'apprentissage avec un superviseur imparfait.

Cependant, ce n'est pas la seule incertitude qui apparaît quand on essaie d'apprendre un prédicteur. En effet, lorsqu'on nous donne un ensemble de données, nous devons faire face à l'incertitude du processus d'échantillonnage : nous n'avons accès qu'à une vue partielle de la distribution latente des données. Un échantillonnage différent des données nous induira à choisir un autre prédicteur. De plus, l'algorithme d'apprentissage peut être intrinsèquement stochastique. Ainsi, lors de l'apprentissage de réseaux neuronaux, en raison de la non-convexité de l'objectif d'apprentissage, la ré-exécution de la descente de gradient stochastique en utilisant une initialisation différente et un brassage différent des données d'apprentissage nous donnera une solution différente qui est souvent comparable en termes de précision tout en apportant de la diversité. Cette incertitude n'est pas intrinsèque à la tâche et survient lors de l'apprentissage du modèle, nous l'appellerons donc *incertitude du modèle*.

Ces deux types d'incertitude, l'ambiguïté des tâches et l'incertitude du modèle, saisissent des informations différentes et complémentaires. Nous proposons de les utiliser explicitement pour construire de nouveaux critères de rejet en essayant de les démêler avant de les fusionner de nouveau de manière adaptée.

Dans cet article, nous nous concentrerons principalement sur la classification binaire à l'aide de réseaux neuronaux. Nous présentons deux contributions principales. Premièrement, dans la Section 3, nous proposons de nouveaux critères de classification avec option de rejet en utilisant des mesures de l'ambiguïté des tâches et de l'incertitude du modèle. Deuxièmement, comme ces critères semblent avoir des performances similaires à celles des méthodes de référence dans nos expériences, dans la Section 4, nous montrons qu'il existe effectivement des informations supplémentaires à exploiter en démêlant les deux types d'incertitude présentés précédemment mais qu'elles peuvent être difficile à saisir sans apprendre un critère ad hoc sur un ensemble de validation.

2 Formulation du problème et état de l'art

La classification avec option de rejet a d'abord été introduite et étudiée dans [4, 5] en utilisant des modèles probabilistes. Des travaux plus récents ont étudié ce problème dans le cadre de la théorie de l'apprentissage statistique en définissant une fonction de risque adaptée. Il est habituellement exprimé pour un coût de rejet donné de λ [14, 6] par

$$R_\lambda(h, r) = \mathbb{E}_{X,Y}\left[\delta_{h(X)\neq Y}(1 - r(X)) + \lambda r(X)\right] \quad (1)$$

et est minimisé pour le classifier de Bayes optimal (h^*, r^*) suivant :

$$h^*(x) = \delta_{\eta(x)\geq\frac{1}{2}} \quad \text{et} \quad r^*(x) = \delta_{|\eta(x)-\frac{1}{2}|<\frac{1}{2}-\lambda}. \quad (2)$$

Il s'agit donc d'un compromis entre le taux d'erreur et le taux de rejet pour ce coût λ et favorise le rejet dans les zones où l'ambiguïté de la tâche est plus grande.

Beaucoup de travaux théoriques partent de cette formulation et se concentrent principalement sur le cas binaire. Par exemple, [14] a étudié le taux de convergence des estimateurs plug-in et des minimiseurs du risque empirique. [1] a proposé une *hinge loss* pour les réjecteurs basés sur la confiance qui peut être mise en œuvre dans la pratique et en propose l'étude théorique. [25] étudie d'autres fonctions objectives convexes tandis que [6] étend cela à des réjecteurs d'une classe de fonctions différente de celle du prédicteur. Une autre approche consiste à apprendre séquentiellement des classifieurs qui peuvent rejeter avec un coût, une tâche appelée apprentissage séquentiel, comme étudié dans [22].

Parallèlement, [8] a utilisé une stratégie basée sur les désaccords pour apprendre un prédicteur parfait dans le cas réalisable, c'est-à-dire en l'absence de bruit et lorsque le classifieur parfait est dans la classe des hypothèses. [23, 24] ont ensuite étudié de telles approches pour le cas agnostique afin d'apprendre ce qu'ils appellent des classifieurs sélectifs point par point, c'est-à-dire que les prédicteurs ont le même taux d'erreur que celui optimal tout en essayant de

maintenir le taux de rejet le plus bas possible. De plus, ils proposent de mesurer la performance de ces modèles, non seulement pour un coût de rejet λ donné, qui pourrait ne pas être quantifiable en pratique, mais sur l'ensemble de la courbe de risque/couverture, ce qui revient à comparer la performance pour tous les choix de λ en même temps [8].
Cependant, ces travaux sont, soit purement théoriques, soit reposent sur une optimisation convexe. Cela peut être un problème pour les réseaux de neurones parce que leur processus d'apprentissage n'est pas encore entièrement compris. Ainsi, de nombreuses études empiriques et heuristiques ont été réalisées autour des critères de rejet et de l'estimation de l'incertitude pour ces modèles. Pour les réseaux neuronaux, les approches basées sur la confiance ont été étudiées depuis longtemps et continuent de l'être jusqu'à récemment dans [17, 16] par exemple. En ce qui concerne les méthodes fondées sur la dispersion, [9] interprète le *dropout* comme une approximation de l'inférence bayésienne et propose d'utiliser la variance du résultat comme mesure d'incertitude. Cependant, [10] montre que ces différentes mesures, basées sur la confiance ou sur la dispersion, peuvent être améliorées en choisissant, pendant la phase d'apprentissage, des modèles mieux adaptés à chaque zone de l'espace d'entrée. Plus largement, dans le contexte de l'apprentissage actif, des mesures basées sur la confiance et sur la dispersion ont été étudiées, [21] en propose une vue d'ensemble.
Les mesures de dispersions que nous introduisons dans ce papier sont basées sur le désaccord entre les modèles d'un ensemble. Cette notion de désaccord est importante dans l'apprentissage et a été exploitée à multiples reprises. En effet, elle a d'abord été utilisée pour obtenir un modèle plus performant que chacun des modèles individuels à travers le *bagging* [3] et le *boosting* [20]. Certains travaux tels que [11] fournissent une étude théorique de ces approaches. La notion de désaccord est également importante dans d'autres paradigmes d'apprentissage tels que l'apprentissage actif [13] ou bien, comme développé précedémment, la classification avec rejet [24].
Enfin, les différentes sources d'incertitudes sont généralement séparées en incertitudes *aléatoriques* et *épistémiques* [7, 15]. La première est définie comme la partie irréductible de l'incertitude, par opposition à la seconde qui peut être réduite en collectant plus de données ou en affinant le modèle. Cependant, comme cela est expliqué en détail dans [7], la distinction précédente a un sens dans un modèle pour lequel il est explicite quelle incertitude peut être réduite. Nous préférons la terminologie de *l'ambiguïté de la tâche* et *l'incertitude du modèle* parce que nous ne considérons pas ici quelle incertitude peut être réduite et comment mais plutôt si elle est intrinsèque à la tâche ou si elle provient du processus d'apprentissage.

3 Deux types d'incertitude

Dans cette section, nous étudions comment, à la fois, l'ambiguïté des tâches et l'incertitude du modèle peuvent être utilisées pour obtenir de nouveaux critères de classification avec option de rejet.

3.1 Incertitude dans le choix du prédicteur

En raison de la stochasticité de l'apprentissage et des processus d'échantillonnage des données, il y a une incertitude dans le choix de l'hypothèse h. Nous pouvons construire un critère basé sur la variabilité de la prédiction due à cette stochasticité.
Comme nous utilisons des classes d'hypothèses paramétrées, l'incertitude dans le choix de $h = h_\theta$ provient en fait d'une incertitude dans le choix des paramètres θ étant donné les données d'apprentissage $\mathcal{D}_{\text{train}}$, c'est-à-dire $\Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]$. Ainsi, une première approche consiste à modéliser complètement cette distribution de probabilité sur les paramètres et une façon de le faire est d'utiliser des approches bayésiennes [18]. Dans ce cas, une distribution *a priori* sur tous les paramètres, $\Pr[\theta = \theta']$, est définie et sa la distribution *a posteriori* $\Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]$ est calculée en appliquant la règle de Bayes.
Cependant, selon le modèle utilisé, il peut être difficile d'établir une distribution *a priori* adaptée, de plus, le calcul de la distribution *a posteriori* peut s'avérer difficile. Il est possible d'utiliser des méthodes d'approximation mais, dans certaines situations, cela peut ne pas être suffisant. Cela est en particulier le cas pour les réseaux de neurones où l'application des méthodes bayésiennes est un domaine de recherche actif [18, 19, 2, 9].
Néanmoins, comme dans le présent article, nous nous intéressons uniquement à la classification avec option de rejet, il n'est pas nécessaire de modéliser cette distribution complexe. Nous pouvons utiliser une approche plus directe pour estimer nos critères en supposant que nous pouvons échantillonner les modèles de notre distribution $\Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]$. On peut y parvenir en simulant les différentes sources d'incertitude du modèle, c'est-à-dire :

- la stochasticité de l'algorithme d'apprentissage en exécutant plusieurs fois l'apprentissage sur les mêmes données
- l'échantillonnage des données d'entraînement en utilisant des techniques de *bootstrap* et de sous-échantillonage pour simuler la stochasticité du processus de génération des données.

Une fois qu'il est supposé que nous pouvons échantillonner à partir de $\Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]$, nous pouvons alors construire des critères basés sur la dispersion. En particulier, nous étudions dans la section suivante un critère basé sur le désaccord entre les prédicteurs de cette distribution.

3.2 Un critère pratique de désaccord

Les critères de désaccord ont été théoriquement étudiés dans le contexte de la classification avec option de rejet pour trouver un prédicteur ayant le même taux d'erreur que le classifieur parfait tout en ayant un taux de rejet faible [8, 23, 24], ainsi que dans le cadre de l'apprentissage actif [13]. Les statégies proposés restent cependant théoriques et

ne sont pas implémentables en pratique. Dans cette sous-section, nous proposons un moyen de les rendre pratiques et de les généraliser.

L'idée générale est de quantifier le désaccord entre les modèles $(h_\theta)_\theta$ d'un ensemble de modèles en regardant uniquement leurs prédictions $\hat{y} = h_\theta(x) \in \mathcal{Y}$. Pour cela, nous définissons la *mesure de désaccord* $\zeta(x)$ comme

$$\zeta(x) = \Pr\left[\hat{Y} = 1 \mid X = x\right] = \mathbb{E}_\theta\left[\delta_{h_\theta(x)=1}\right],$$

étant donné une mesure de probabilité sur l'ensemble des paramètres θ. Notez que cette quantité est différente de la fonction de régression $\eta(x)$ qui mesure l'ambiguïté de la tâche elle-même, $\Pr[Y = 1 \mid X = x]$, indépendamment des modèles. De plus, cette quantité n'est pas à proprement parler une probabilité de désaccord, mais elle capture cette information. En effet, le désaccord est maximal quand $\zeta(x)$ vaut 0.5 et minimal quand il vaut 0 ou 1.

Si nous modélisons complètement l'incertitude dans les paramètres $\Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]$, il est alors possible de calculer le désaccord entre ces hypothèses en utilisant

$$\hat{\zeta}(x) = \int_{\theta'} \delta_{h_{\theta'}(x)=1} \Pr[\theta = \theta' \mid \mathcal{D}_{\text{train}}]\, d\theta'. \quad (3)$$

Cependant, si nous ne nous intéressons qu'à $\zeta(x)$, nous n'avons, en général, pas besoin de modéliser complètement la distribution sur les paramètres, ce qui peut être une tâche complexe comme le montre la sous-section précédente.

Au lieu de cela, parce que $\hat{y} \in \mathcal{Y}$ est discret, si on dispose de C échantillons, $\theta_1, \ldots, \theta_C$, nous pouvons estimer directement la distribution $\Pr\left[\hat{Y} = 1 | X = x\right]$ avec une approche fréquentiste :

$$\hat{\zeta}_{\text{freq}}(x) = \frac{1}{C}\sum_{i=1}^{C} \delta_{h_{\theta_i}(x)=1}. \quad (4)$$

Néanmoins, bien que cet estimateur soit non biaisé, il pourrait nécessiter beaucoup d'échantillons, soit un grand C, pour réduire sa variance.

En général, la plupart des modèles h_θ peuvent être décomposés en utilisant une fonction paramétrée $f_\theta : \mathcal{X} \to \mathbb{R}$ sur lequel est appliquée une fonction de décision $\delta_Z : \mathbb{R} \to \{0, 1\}$. Typiquement, $\delta_Z(z) = \delta_{z\geq 0}$. Lors de l'utilisation d'une fonction de coût logistique, z est appelé logit et nous pouvons même décomposer h un peu plus en appliquant, par-dessus f_θ, la fonction sigmoid $\sigma : \mathbb{R} \to [0, 1]$ pour transformer z en probabilité avant de prendre la décision qui devient $\delta_P(p) = \delta_{p\geq \frac{1}{2}}$. Ces décompositions peuvent être résumées comme suit :

$$x \xrightarrow{h_\theta} \hat{y} \;\Leftrightarrow\; x \xrightarrow{f_\theta} z \xrightarrow{\delta_Z} \hat{y} \;\Leftrightarrow\; x \xrightarrow{f_\theta} z \xrightarrow{\sigma} p \xrightarrow{\delta_P} \hat{y}$$

Sur la base de la décomposition précédente, nous pouvons en fait utiliser une approche intermédiaire entre les deux extrêmes que sont les Équations (3) et (4). En effet, parce que $\hat{y}$ est une fonction (non paramétrique) de z et p, il suffit de modéliser l'incertitude dans l'espace des logits ou de la probabilité. A partir de ces distributions, nous pouvons ensuite dériver la mesure de désaccord en utilisant

$$\hat{\zeta}(x) = \int_z \delta_{z\geq 0} \Pr[Z = z \mid X = x]\, dz = 1 - F_x^Z(0) \quad (5)$$

et

$$\hat{\zeta}(x) = \int_p \delta_{p\geq \frac{1}{2}} \Pr[P = p \mid X = x]\, dp = 1 - F_x^P(\frac{1}{2}) \quad (6)$$

où F_x^Z et F_x^P sont respectivement la fonction de répartition de la probabilité conditionnelle dans l'espace des logits et des probabilités.

Cette méthode permet d'introduire des hypothèses et des connaissances à priori qui sont plus faciles à tester dans la pratique que la modélisation complète de la distribution des paramètres. En même temps, si ces hypothèses sont vérifiées, l'estimateur résultant convergerait plus rapidement que $\hat{\zeta}_{\text{freq}}(x)$, ce qui signifie que nous pourrions choisir un C plus petit, rendant cette approche plus pratique. Finalement, cette formalisation nous permet de mieux comprendre comment fusionner les différentes statistiques et moments, comme le montre la section suivante.

3.3 Choix de la distribution

Nous étudions maintenant quels sont les choix de distributions appropriées dans les espaces des logits et des probabilités.

En supposant qu'on nous donne plusieurs logits, $z_1, \ldots, z_C$, pour une entrée x, un choix courant est d'utiliser une loi normale pour modéliser leur distribution. Dans ce cas, nous pouvons ajuster les paramètres de la distribution en utilisant les estimateurs de maximum de vraisemblance habituels $\hat{\mu}_z = \frac{1}{C}\sum_{i=1}^{C} z_i$ et $\hat{\sigma}_z^2 = \frac{1}{C}\sum_{i=1}^{C}(z_i - \hat{\mu}_z)^2$. L'estimateur de l'Équation (5) devient alors

$$\hat{\zeta}_{\text{norm}}(x) = 1 - \phi\left(-\frac{\hat{\mu}_z}{\hat{\sigma}_z}\right),$$

où ϕ est la fonction de répartition de la distribution normale standard. Fait intéressant, ce critère est une fonction bijective de $\frac{\hat{\mu}_z}{\hat{\sigma}_z}$.

Alternativement, si nous considérons l'espace des distributions binaires p, un choix naturel est la loi bêta. Dans ce cas, il n'existe pas de forme close pour les estimateurs du maximum de vraisemblance de ses paramètres α et β. Il faut soit s'appuyer sur un algorithme itératif, soit utiliser la méthode des moments. Dans ce dernier cas, les estimateurs de la méthode des estimateurs sont égaux à $\hat{\alpha} = \hat{\mu}_p\left(\frac{\hat{\mu}_p(1-\hat{\mu}_p)}{\hat{v}_p} - 1\right)$ et $\hat{\beta} = (1 - \hat{\mu}_p)\left(\frac{\hat{\mu}_p(1-\hat{\mu}_p)}{\hat{v}_p} - 1\right)$ avec $\hat{\mu}_p = \frac{1}{C}\sum_{i=1}^{C} p_i$ et $\hat{v}_p = \frac{1}{C}\sum_{i=1}^{C}(p_i - \hat{\mu}_p)^2$. L'estimateur de l'Équation (6) est alors égal à

$$\hat{\zeta}_{\text{beta}}(x) = 1 - I_{\frac{1}{2}}(\hat{\alpha}, \hat{\beta}),$$

où I_x est la fonction bêta incomplète régularisée.

Maintenant que nous disposons d'une mesure pratique de désaccord, nous pouvons établir un critère de classification avec une option de rejet telle que

$$c_{\text{disagree}}(x) = \max(\hat{\zeta}(x), 1 - \hat{\zeta}(x)). \quad (7)$$

Cependant, il ne s'agit là que d'une mesure de la dispersion des prédictions de l'ensemble, l'incorporation d'une mesure de l'ambiguïté de la tâche pourrait conduire à de meilleurs critères.

3.4 Critère de fusion

Si l'on examine la fonction de risque de la classification avec option de rejet de l'Équation (1), cette quantité est minimisée par le réjecteur optimal de Bayes de l'Équation (2). En estimant la fonction de régression $\eta(x)$, nous pouvons construire une règle *plug-in* qui nous permettrait d'effectuer un rejet basé sur notre estimateur $\hat{\eta}(x)$. Le taux de convergence théorique de cet estimateur *plug-in* a été étudié dans [14] où il a été démontré qu'il dépend de la qualité de l'estimateur $\hat{\eta}(x)$ et de la structure et du niveau de l'ambiguité $\eta(x)$. La construction d'un tel estimateur peut se faire en optimisant une fonction de coût *strictly proper* [12]. Il s'avère que la fonction de coût logistique est en fait *strictly proper*, un telle approche a par ailleurs été appliquée aux réseaux neuronaux dans [16].
Cependant, cette méthode ne tient pas compte de la variabilité de l'estimateur qui change en fonction des zones de l'espace d'entrée. Cette variabilité pourrait donner lieu à une prédiction différente $\hat{y}$. En utilisant le critère de désaccord de la sous-section précédente, nous pouvons construire un nouveau critère en marginalisant notre incertitude dans le choix de la prédiction :

$$c_{\text{fusion}}(x) = \Pr\left[\hat{Y} = 0 \mid X = x\right](1 - \hat{\eta}(x)) + \Pr\left[\hat{Y} = 1 \mid X = x\right]\hat{\eta}(x).$$

S'il n'y a pas de désaccord, c'est-à-dire si $\zeta(x) \in \{0, 1\}$, ce critère est simplement l'ambiguïté estimée de la prévision : $\max(\hat{\eta}(x), 1 - \hat{\eta}(x))$. De plus, pour une valeur égale de $\hat{\eta}(x)$, ce critère rejettera d'abord les zones de l'espace d'entrée où le désaccord est le plus fort. Cette quantité a donc des propriétés intéressantes.
En injectant notre estimateur de $\zeta(x)$ dans l'équation précédente, le critère devient

$$c_{\text{fusion}}(x) = (1 - \hat{\zeta}(x))(1 - \hat{\eta}(x)) + \hat{\zeta}(x)\hat{\eta}(x). \quad (8)$$

3.5 Expériences synthétiques

Afin de comparer les différents critères de rejet, nous utilisons des jeux de données synthétiques où nous pouvons contrôler la quantité d'ambiguïté $\eta(x)$ et la densité des données $p(x)$. Nous utilisons trois jeux de données différents illustrés dans la Figure 1 :

- Fig.1a : absence d'ambiguïté, i.e. $\eta(x) \in \{0, 1\}$, mais il existe deux zones de densités uniformes différentes ;
- Fig.1b : x est distribué uniformément mais il y a des zones d'ambiguïté de quantité différente ;
- Fig.1c : mélange des deux scénarios précédents.

Ces jeux de données permettent de comprendre finement comment les prédictions du modèle et les critères de rejet se comportent sous différentes contraintes d'incertitude.

La mesure de performance que nous utilisons est la courbe RC [8] et, en particulier, l'aire sous cette courbe telle que définie dans [10] que nous désignons RC-AUC. La courbe de risque/couverture (RC) est définie comme suit

$$R(h, r) = \mathbb{E}_{X,Y}\left[\delta_{h(X)\neq Y}\frac{1 - r(X)}{\phi(h, r)}\right],$$
$$\phi(h, r) = 1 - \mathbb{E}_X\left[r(X)\right].$$

Le risque R quantifie le taux d'erreur des échantillons qui ne sont pas rejetés. La couverture ϕ mesure le taux d'acceptation. Plus cette quantité est faible, meilleur est le compromis entre précision et taux de rejet.
Toutes les fonctions de rejet que nous considérons effectuent un seuillage se basant sur un critère c :

$$r(x) = \delta_{c(x)\geq\tau}. \quad (9)$$

Nous utilisons comme références les critères suivants :

- pour les approches basées sur la confiance : moyenne de $|\mu_p - \frac{1}{2}|$ dans l'espace de probabilité, notée "prob. mean", et moyenne de $|\mu_z|$ dans l'espace logit, notée "logit mean" [1]
- pour les approches basées sur la dispersion : la variance de $|\mu_p - \frac{1}{2}|$ notée "var. prob." et de $|\mu_z|$ notée "var. logit" et la moyenne de la divergence de Kullback-Leibler (KL) des prévisions des modèles individuels de l'ensemble comparé à la prévision moyenne telle que définie dans [21]

Notez que pour les critères basés sur la dispersion, c'est leur opposé qui est seuillé. Nous intégrons en outre le classifieur optimal de Bayes de l'Équation (2) pour fournir la plus faible valeur de la RC-AUC réalisable.
Nous comparons d'abord les différents critères basés sur les mesures de désaccord des Équations (7) et (8) dans la colonne de gauche de la Figure 2. Avec suffisamment de données, les critères de désaccord convergent vers le critère optimal en l'absence d'ambiguïté alors qu'en présence d'ambiguïté, ce n'est pas le cas. Cela est attendu car ces critères ne prennent pas en compte cette partie de l'incertitude, mais seulement le désaccord entre les modèles de l'ensemble. L'utilisation du critère de fusion pour incorporer l'ambiguïté à ces critères les améliore et les fait converger vers la performance du critère optimal. Parce que nous utilisons un ensemble de taille 1000, les approches fréquentistes peuvent être considérées comme les meilleurs estimateurs des critères de désaccord et de fusion. Les critères basés sur la loi bêta sont toujours assez proches et, par conséquent, cette dernière peut être considérée comme un bon choix pour modéliser la distribution des probabilités produites par l'ensemble. A contrario, les critères basés sur la loi normale ont une performance médiocre en l'absence d'ambiguïté tout en étant au même niveau ou légèrement meilleurs autrement. Cette loi ne semble donc pas adaptée pour modéliser la distribution dans l'espace des logits.

1. ici, parce que nous considérons des tâches de classification binaire, les autres critères basés sur ces moyennes, par exemple, l'entropie etc, sont équivalents aux critères que nous utilisons

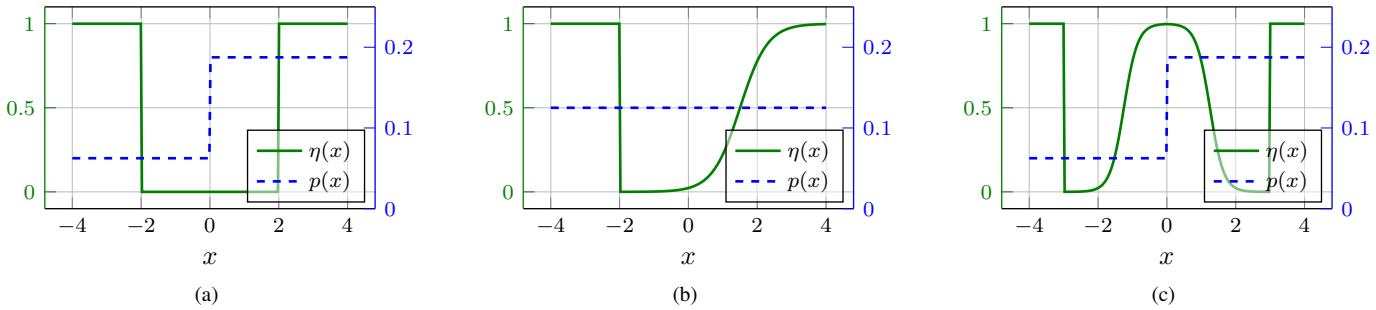


FIGURE 1 – Jeux de données synthétiques avec un degré variable d'ambiguïté et de densité des données. La distribution de densité de probabilité de x est indiquée en pointiés bleus tandis que la fonction de régression $\eta(x)$ est indiquée en vert.

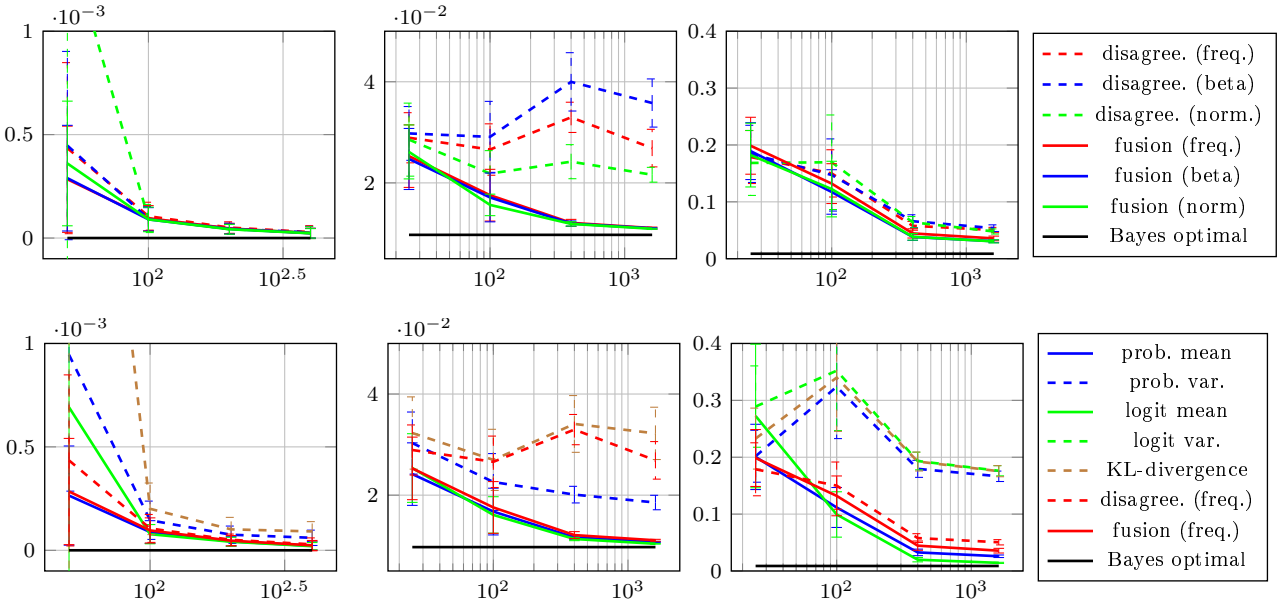


FIGURE 2 – Aire sous la courbe des courbes risk/couverture (RC-AUC) pour comparer les différents critères en fonction de la taille du jeu d'apprentissage. Les colonnes correspondent aux trois jeux de données synthétiques de la Figure 1. La première ligne compare les critères de désaccord et les critères de fusion entre eux. La seconde ligne les compare aux critères de base.

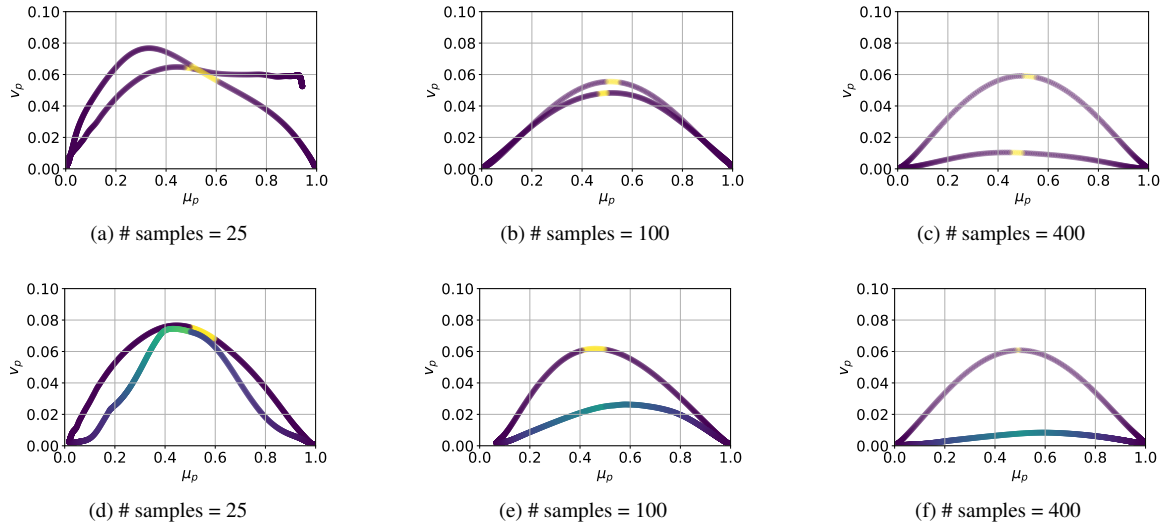
Lorsque ces critères sont comparés à ceux de référence, les critères de fusion donnent des résultats proches de la moyenne de la probabilité et se situent toujours à l'intérieur des fluctuations statistiques comme le montre la colonne de droite de la Figure 2. De plus, il est surprenant de constater que les autres critères basés sur la dispersion peuvent avoir de mauvais résultats et, dans certains cas, ils peuvent même ne pas apparaître sur le graphique.

Deux conclusions principales se dégagent de ces expériences. Premièrement, le bon comportement du critère fondé sur la moyenne de probabilité indique qu'il semble saisir une partie de l'incertitude du modèle, ce qui implique qu'il s'agit d'un estimateur biaisé en ce sens de $\eta(x)$. Par exemple, en l'absence d'ambiguïté, il est toujours performant alors que l'ambiguïté satisfait $\eta(x) \in \{0, 1\}$ et n'apporte donc aucune information sur l'incertitude dans ce cas. Deuxièmement, le critère de fusion proposé ne semble pas tenir compte de l'incertitude supplémentaire, du moins dans le cadre de ces expériences synthétiques. Toutefois, dans la section suivante, nous analysons plus en détail ces jeux de données contrôlés afin de comprendre s'il existe effectivement des informations supplémentaires qui peuvent être exploitées et comment y parvenir.

4 Étude approfondie des rejeteurs

4.1 Visualisation des frontières de décision

Les critères précédemment étudiés peuvent être considérés comme une simple décision prise par seuillage comme explicité dans l'Équation (9). La plupart d'entre eux définissent des frontières de décision, soit sur la moyenne et la variance dans l'espace de probabilité, (μ_p, v_p), soit



(a) # samples = 25 (b) # samples = 100 (c) # samples = 400

(d) # samples = 25 (e) # samples = 100 (f) # samples = 400

FIGURE 3 – La moyenne et la variance des prédictions dans l'espace de probabilité, (μ_p, v_p), pour chaque point de l'espace d'entrée, c'est-à-dire $-4 < x < 4$. La couleur de la courbe indique le taux d'erreur pour ces valeurs, le jaune correspondant à un taux d'erreur élevé et le mauve à un taux faible, le blue étant une valeur intermédiaire. Les première et deuxième lignes correspondent respectivement au premier et au deuxième jeu de données synthétique.

sur la moyenne et l'écart-type dans l'espace logit, (μ_z, σ_z). Afin de visualiser la complexité de la frontière de décision idéale, la Figure 3 montre les valeurs de (μ_p, v_p) pour x variant sur l'espace d'entrée, de -4 à 4, pour les deux premiers jeux de données synthétiques. À chacun de ces points est associé la probabilité d'erreur en couleur.

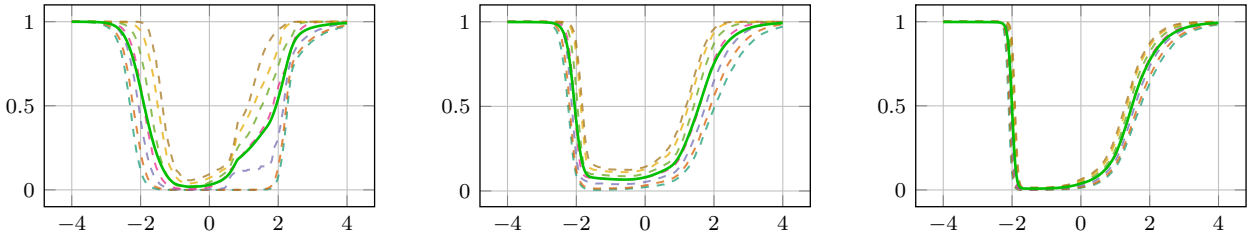
Ces figures montrent que cette frontière de décision idéale peut en fait être assez complexe et varie en fonction de la taille du jeu d'apprentissage. De plus, la variance de la zone avec moins d'incertitude, c'est-à-dire pour x autour de -2, augmente avec le nombre d'échantillons d'apprentissage et cette zone finit par être celle avec la variance la plus élevée. Au premier abord, cela semble plutôt contre-intuitif. En effet, si l'on s'attend à ce que la variance donne directement l'information de l'incertitude du modèle, cette quantité devrait toujours diminuer au fur et à mesure que la taille des données d'entrainement augmente. De plus, il devrait être plus faible dans la zone de non-ambiguïté parce que le modèle devrait la capturer plus rapidement.

Si l'on examine de près la distribution des fonctions de prédiction des modèles dans l'ensemble de la Figure 4, cela est en fait logique. Dans les zones où l'ambiguïté est faible, lorsqu'on dispose de suffisamment de données d'apprentissage, l'incertitude de la fonction de prévision est faible et nous savons presque exactement où se produit le passage d'une classe à une autre. Cependant, comme tous les modèles de l'ensemble ont compris qu'il n'y a pas d'ambiguïté, chacun d'entre eux prédit ou bien 0, ou bien 1, avec une "confiance" élevée, ce qui entraîne une variance élevée de cette valeur de confiance à ce point précis tout en ayant un faible risque d'erreur. A contrario, dans les zones ambigües, la variance est plus faible car chaque modèle a connaissance de cette ambiguité mais est incertain sur l'endroit exact où la transition de la classe 0 à la classe 1 a lieu. La fonction de décision finale semble plus incertaine mais la variance calculée en un point reste relativement faible comparé à la zone non-ambigüe.

Les Figure 5 et Figure 6 montrent les frontières de décision des critères de rejet utilisés dans la section précédente, respectivement, dans l'espace de probabilité et l'espace logit. Les critères de désaccord et de fusion que nous proposons tiennent compte à la fois de la moyenne et de la variance dans ces espaces et sont donc plus susceptibles de rejeter des zones de forte variance même pour une moyenne constante ou le contraire. Cependant, ces graphiques mettent en lumière les limites des différents critères de base et des critères que nous proposons. En effet, aucun d'entre eux n'a une forme adaptée pour effectuer un rejet efficace sur les jeux de données synthétiques. Cela est attendu, car plus la variance est élevée, plus il est probable que nous rejettions. Mais cette hypothèse n'est pas la bonne ici, ce qui souligne la complexité de l'élaboration d'un critère théorique. Cependant, comme on peut le remarquer dans la Figure 3, les zones de taux d'erreur différents peuvent toujours être séparées dans l'espace de (μ_p, v_p). Nous pouvons donc essayer d'apprendre de manière supervisée à mieux rejeter.

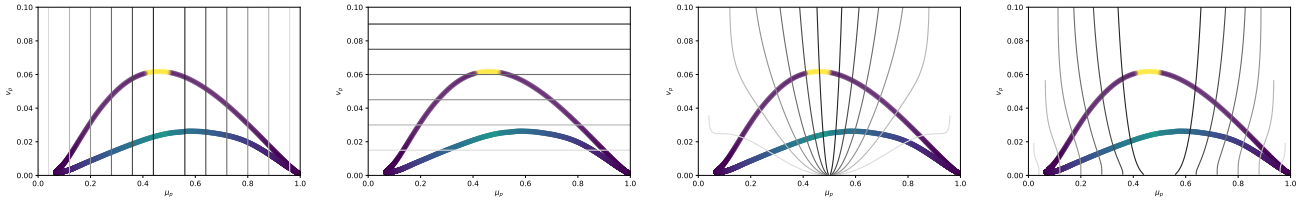
4.2 Apprentissage supervisé du réjecteur

Étant donnée que la frontière de décision du réjecteur est une fonction complexe, comme nous l'avons vu dans la sous-section précédente, nous proposons d'essayer de l'apprendre à partir des données en utilisant un jeu de données de calibration. Pour ce faire, nous devons revenir à la fonction de risque de la classification avec option de rejet de l'Équation (1). Dans notre cas, h est déjà appris et il suf-



(a) # samples = 25 (b) # samples = 100 (c) # samples = 400

FIGURE 4 – Diagrammes en centiles des distributions binaires prédites par les modèles de l'ensemble pour plusieurs tailles du jeu d'apprentissage. Les centiles représentés en pointillés sont : 5%, 10%, 25%, 50%, 75%, 90% et 95%. La moyenne de ces prédictions est indiquée en vert.



(a) moyenne des probabilités (b) variance des probabilités (c) désaccord (bêta) (d) fusion (bêta)

FIGURE 5 – Frontières de décision des différents critères dans l'espace (μ_p, v_p) de l'espace de probabilité.

fit donc d'optimiser le risque précédent par rapport à r. Ce scénario a été étudié dans le contexte de l'apprentissage séquentiel [22]. Cela revient simplement à apprendre une tâche de classification binaire pondérée

$$R_\lambda^{\text{rej}}(h, r) = \mathbb{E}_{X,E}\left[w_E \delta_{r(X)\neq E}\right], \tag{10}$$

où E est la variable aléatoire correspondant à une erreur de prédiction, $E = \delta_{h(X)\neq Y}$, et les poids sont égaux à $w_0 = \lambda$ et $w_1 = 1 - \lambda$.

Le réjecteur de Bayes optimal est dans ce cas égal à

$$r^*(x) = \delta_{\eta_E(x)>\lambda} \tag{11}$$

où $\eta_E(x)$ est la fonction de régression de l'Équation (10), c'est-à-dire $\eta_E(x) = \Pr\left[Y \neq h(X) \mid X = x\right]$.

Nous pouvons alors apprendre n'importe quel modèle de classification habituel pour effectuer cette tâche. En faisant varier la valeur de λ, on peut alors reconstituer la courbe RC en réapprenant le réjecteur adapté à chaque fois.

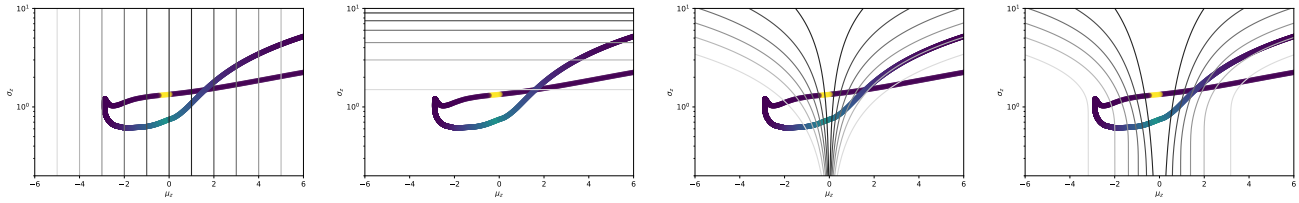
Notre but ici n'est pas de construire le meilleur modèle de rejet mais plutôt de montrer qu'en changeant notre espace d'entrée pour le rejet de $x \in \mathcal{X}$ à $(\mu_p, v_p) \in \mathbb{R}^2$, nous conservons encore suffisamment d'informations discriminantes pour effectuer le rejet. Nous ne prétendons pas qu'il s'agit du meilleur espace pour accomplir cette tâche mais qu'il est suffisant pour apprendre de meilleurs critères de rejet que les critères précédents.

Nous utilisons des classifieurs polynomiaux de degré 3 entrainés avec une fonction de coût logistique. Nous utilisons beaucoup de données d'étalonnage, soit 1000 échantillons, pour apprendre le réjecteur. Ce chiffre n'est pas réaliste dans un scénario réel où nous préférerions utiliser ces données pour améliorer notre prédicteur. Cependant, notre but ici est de comprendre à quel point un dispositif de rejet peut devenir meilleur si nous lui permettons d'être plus complexe et d'être dépendant de la tâche. La Figure 7 montre la courbe RC-AUC du rejecteur par rapport aux critères précedents et aux réjecteurs de Bayes des Équations (11) et (2) sur le deuxième jeu de données synthétique. La fonction de rejet apprise est en effet bien meilleure que les critères précédents. Cela montre la richesse de l'information contenue dans le plan (μ_p, v_p).

5 Conclusion et travaux futurs

Les critères classiques, tels que la valeur maximale de la probabilité prédite par un réseau neuronal, fonctionnent bien dans la pratique, cependant ils n'utilisent pas toute l'information d'incertitude disponible. L'analyse des prédictions d'un ensemble met en lumière certains comportements contre-intuitifs qui soulignent notre mauvaise compréhension de l'information d'incertitude capturée par ces modèles. Trouver un bon critère de rejet qui tire parti de toute cette information et qui généralise à différentes tâches est un problème difficile.

Une suite naturelle de ces travaux est d'effectuer des analyses sur des données réelles afin de vérifier que les comportements présentés ici se produisent également dans des jeux de données usuels. De plus, ce travail ouvre la voie à l'élaboration d'un meilleur critère de rejet. Une piste de recherche en ce sens consiste à trouver un espace de représentation plus adapté pour l'apprentissage supervisé des réjeteurs afin de permettre à ces derniers d'être calibrés en



(a) moyenne des logits (b) écart-type des logits (c) désaccord (normale) (d) fusion (normale)

FIGURE 6 – Frontières de décision des différents critères dans l'espace (μ_z, σ_z) de l'espace des logits.

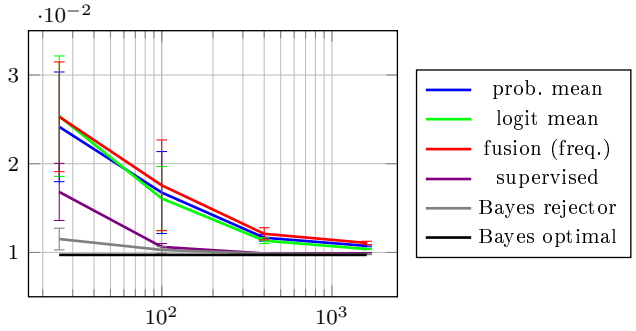


FIGURE 7 – Aire sous la courbe des courbes RC (RC-AUC) comparant le réjecteur supervisé aux différents autres critères avec une taille du jeu d'apprentissage variable sur le deuxième jeu de données synthétique.

utilisant beaucoup moins de données.

Remerciements

Ce travail a été partiellement financé par le projet ANR WeedElec.

Références

[1] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 2008.

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *ICML*, 2015.

[3] L. Breiman. Bagging predictors. *Mach. Learn.*, 1996.

[4] C. K. Chow. An optimum character recognition system using decision functions. *IRE T. Elec. Comp.*, 1957.

[5] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 1970.

[6] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *ALT*, 2016.

[7] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic ? Does it matter ? *Struct. Saf.*, 2009.

[8] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 2010.

[9] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation : representing model uncertainty in Deep Learning. In *ICML*, 2016.

[10] Y. Geifman, G. Uziel, and R. El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*, 2019.

[11] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. Risk bounds for the majority vote : from a PAC-Bayesian analysis to a learning algorithm. *J. Mach. Learn. Res.*, 2015.

[12] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 2007.

[13] S. Hanneke et al. Theory of disagreement-based active learning. *Found. and Trends® in Mach. Learn.*, 2014.

[14] R. Herbei and M. H. Wegkamp. Classification with reject option. *Can. J. Stat.*, 2006.

[15] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian Deep Learning for Computer Vision ? In *NeurIPS*, 2017.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[17] A. Mandelbaum and D. Weinshall. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv :1709.09844*, 2017.

[18] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[19] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[20] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 1999.

[21] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.

[22] K. Trapeznikov and V. Saligrama. Supervised sequential classification under budget constraints. In *AISTATS*, 2013.

[23] Y. Wiener and R. El-Yaniv. Agnostic selective classification. In *NeurIPS*, 2011.

[24] Y. Wiener and R. El-Yaniv. Agnostic pointwise-competitive selective classification. *J. Artif. Intell. Res.*, 2015.

[25] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.*, 2010.