Classifieur minimax discret randomisé pour la détection de classes rares et la présence de probabilités a priori imprécises

Cyprien GILET[1] Marie GUYOMARD[2] Sébastien DESTERCKE[1] Lionel FILLATRE[2]

[1]Université de Technologie de Compiègne, Laboratoire Heudiasyc, France

[2]Université Côte d'Azur, Laboratoire I3S, France

Résumé – Cet article propose un nouveau classifieur quasi-Bayésien égalisateur randomisé pour détecter des classes rares et qui est robuste face aux changements de probabilités a priori. Ce classifieur peut être appliqué sur tout type de données discrètes ou discrétisées. En relaxant l'optimalité au sens de Bayes et en autorisant des décisions randomisées, nous proposons un algorithme qui permet de calculer un classifieur qui égalise tous les risques conditionnels par classe. Le classifieur obtenu est ainsi minimax car les maximums des risques conditionnels par classe sont minimisés.

Abstract – This paper proposes a new randomized equalizer quasi-Bayesian classifier to detect rare classes that is robust to changes in prior probabilities. This classifier can be applied to any discrete or discretized data. By relaxing optimality in the sense of Bayes and allowing randomized decisions, we propose an algorithm that computes a classifier equalizing all conditional risks. The resulting classifier is thus minimax since the maximum of the risks per class is minimized.

1 Introduction et Motivations

Motivations. Nous souhaitons utiliser la classification supervisée pour détecter des classes rares (tri de déchets recyclables, détection d'anomalies, de pannes...) à partir de données contenant à la fois des variables catégorielles et numériques, ainsi qu'en présence de probabilités à priori imprécises. Les classes rares apparaissent peu souvent et elles sont sous-représentées dans les ensembles d'apprentissage alors qu'elles sont parfois les classes les plus pertinentes à détecter (par exemple, une panne dangereuse). L'approche minimax construit un classifieur insensible à la probabilité d'apparition initiale des classes. Puisqu'il privilégie les classes rares, il peut éventuellement dégrader la bonne classification des classes fréquentes.

Contexte et notations. Définissons $K \geq 2$ le nombre de classes, $\mathcal{Y} := \{1, \ldots, K\}$ l'ensemble des classes à prédire, $\mathcal{X}$ l'espace sur lequel les variables descriptives sont définies, n le nombre d'observations composant la base d'apprentissage et n_k le nombre d'observations dans chaque classe $k \in \mathcal{Y}$. On note Y_i la variable aléatoire caractérisant la classe de l'observation i, et $X_i = [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ le vecteur aléatoire regroupant l'ensemble des d variables descriptives associées à l'observation i. On considère de plus une fonction de perte $L : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ qui mesure la perte $L(k, l) = L_{kl}$ lorsque le classifieur décide la classe l alors que la vraie classe est k. Définissons enfin $\Delta = \{\delta : \mathcal{X} \to \mathcal{Y}\}$ l'ensemble des classifieurs. Cet ensemble Δ contient les classifieurs déterministes ainsi que les classifieurs randomisés. Contrairement à un classifieur déterministe, un classifieur randomisé attribuera une classe $k \in \mathcal{Y}$ aléatoirement pour chaque profil $X_i \in \mathcal{X}$ à partir de scores de probabilités estimés sur la base d'apprentissage. Pour la suite, l'ensemble des classifieurs déterministes sera noté $\mathcal{D} \subset \Delta$. Enfin, nous noterons $\mathbb{S} := \{\pi \in [0, 1]^K : \sum_{k \in \mathcal{Y}} \pi_k = 1\}$ le simplexe de dimension K (l'ensemble des probabilités a priori).

Risques d'erreurs moyen et par classe. À partir d'un ensemble fini d'observations étiquetées $\mathcal{S} = \{(Y_i, X_i), i \in \mathcal{I}\}$, l'apprentissage d'un classifieur supervisé $\delta \in \Delta$ consiste généralement à minimiser le risque empirique moyen d'erreurs de classification $\hat{r}(\delta)$ définit par

$$\hat{r}(\delta) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \delta(X_i)) = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k(\delta). \quad (1)$$

Dans l'équation (1), $\hat{\pi}_k = n_k/n$ correspond à la proportion d'observations appartenant à la classe k et

$$\hat{R}_k(\delta) = \sum_{l \in \mathcal{Y}} L_{kl} \hat{\mathbb{P}}(\delta(X_i) = l \mid Y_i = k) \quad (2)$$

correspond au risque d'erreur conditionnel associé à la classe k, avec enfin pour chaque classe $l \in \mathcal{Y}$ et chaque classe $k \in \mathcal{Y}$, $\hat{\mathbb{P}}(\delta(X_i) = l \mid Y_i = k) = \frac{1}{n_k} \sum_{i:Y_i=k} \mathbb{1}_{\{\delta(X_i)=l\}}$.

De plus, comme précisé dans [1, 7, 2, 3], si les proportions par classe (ou probabilités a priori) viennent à évoluer entre la base d'apprentissage et les observations de test, alors le risque d'erreur moyen associé à ces nouvelles probabilités a priori $\pi = [\pi_1, \ldots, \pi_K] \in \mathbb{S}$ satisfera

$$\hat{r}(\pi, \delta) = \sum_{k \in \mathcal{Y}} \pi_k \hat{R}_k(\delta). \quad (3)$$

Autrement dit, le risque d'erreur moyen évolue linéairement entre $[\min_{k \in \mathcal{Y}} \hat{R}_k(\delta), \max_{k \in \mathcal{Y}} \hat{R}_k(\delta)]$ lorsque les probabilités a priori évoluent sur le simplexe. En présence de classes difficiles à prédire, les risques d'erreurs conditionnels associés $\hat{R}_k(\delta)$ sont très élevés par rapport aux autres classes plus facilement identifiables, ce qui entraînera ainsi une grande évolution du risque d'erreur moyen en cas d'évolution des probabilités a priori pour les observations test.

Classifieur minimax déterministe pour données discrètes. La théorie de la décision a démontré dans [1, 2, 7] qu'une solution idéale pour obtenir un classifieur optimal face aux problèmes de classes difficiles à prédire et de probabilités a priori incertaines est de calculer un classifieur minimax

$$\delta^M = \operatorname*{argmin}_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \hat{r}(\pi, \delta) = \operatorname*{argmin}_{\delta \in \Delta} \max_{k \in \mathcal{Y}} \hat{R}_k(\delta). \quad (4)$$

Le classifieur minimax correspond généralement à un classifieur de Bayes pour lequel les risques par classe sont tous égaux [1, 2, 7]. Toutes les classes sont prédites équitablement.

Nos précédents travaux dans [3] avaient pour objectif de calculer un classifieur minimax déterministe en présence de variables descriptives numériques et catégorielles. Afin de pouvoir obtenir une estimation précise et analytique du risque de Bayes sur le simplexe $\mathbb{S}$, nous partitionnons préalablement l'espace des variables descriptives $\mathcal{X}$ en T régions disjointes $\{\Omega_1, \ldots, \Omega_T\}$ telles que $\cup_{t=1}^{T} \Omega_t = \mathcal{X}$. Ceci définit une fonction de cartographie $\phi : \mathcal{X} \mapsto \mathcal{T} := \{1, \ldots, T\}$ telle que $\phi(X_i) = t$ si et seulement si $X_i \in \Omega_t$. Par exemple, l'ensemble des profils discrets $\mathcal{T} = \phi(\mathcal{X})$ peut correspondre aux feuilles d'un arbre de décision ou aux centroïdes issus d'un partitionnement par l'algorithme Kmeans.

Considérons $\mathcal{I}_k = \{i \in \mathcal{I} : Y_i = k\}$ les observations d'apprentissage issues de la classe $k \in \mathcal{Y}$ et $n_k = |\mathcal{I}_k|$ l'effectif dans $\mathcal{I}_k$. À partir de l'ensemble $\mathcal{T} = \phi(\mathcal{X})$ nous pouvons estimer les probabilités $\hat{p}_{kt}$ d'observer le profil discret t sachant que la vraie classe est k, quelque soit $t \in \mathcal{T}$ et quelque soit $k \in \mathcal{Y}$, telles que

$$\hat{p}_{kt} := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \mathbb{1}_{\{\phi(X_i)=t\}}. \tag{5}$$

Pour la suite, ces probabilités $\hat{p}_{kt}$ estimées sur la base d'apprentissage seront considérées comme fixes. Dans [3] et à partir de l'équation (5), le classifier de Bayes déterministe $\delta_\pi^B := \operatorname{argmin}_{\delta \in \mathcal{D}} \hat{r}(\pi, \delta)$ pour données discrétisées et calculé analytiquement quelque soit $\pi \in \mathbb{S}$:

$$\delta_\pi^B : X_i \mapsto \operatorname{argmin}_{l \in \mathcal{Y}} f_l(\pi, X_i), \quad \text{avec } \forall l \in \mathcal{Y}, \tag{6}$$

$$f_l(\pi, X_i) := \textstyle\sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} L_{kl}\, \pi_k\, \hat{p}_{kt}\, \mathbb{1}_{\{\phi(X_i)=t\}}. \tag{7}$$

De plus, son risque $V_B(\pi) := \min_{\delta \in \mathcal{D}} \hat{r}(\pi, \delta) = \hat{r}\left(\delta_\pi^B\right)$ est analytiquement donné par

$$V_B(\pi) = \sum_{k \in \mathcal{Y}} \pi_k \underbrace{\sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{Y}} L_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\{\xi_{lt} = \min_{q \in \mathcal{Y}} \xi_{qt}\}}}_{\hat{R}_k(\delta_\pi^B)}, \tag{8}$$

où quelque soit $(l, t) \in \mathcal{Y} \times \mathcal{T}$, $\xi_{lt} = \sum_{j \in \mathcal{Y}} L_{jl}\, \pi_j\, \hat{p}_{jt}$. Il est par ailleurs établi dans [3] que le risque de Bayes $V_B(\pi)$ est une fonction concave affine par morceaux avec un nombre fini de faces sur le simplexe $\mathbb{S}$. Une illustration de cette propriété est disponible dans la Figure 1. Finalement, notre classifieur minimax déterministe calculé dans [3], noté $\delta_{\bar{\pi}}^B$, correspond au classifieur de Bayes (6) associé aux probabilités a priori $\bar{\pi} = \operatorname{argmax}_{\pi \in \mathbb{S}} V_B(\pi)$. Ce classifieur représente le classifieur déterministe pour lequel les risques par classe sont les plus équilibrés et qui satisfait $\max_{k \in \mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B) = \min_{\delta \in \mathcal{D}} \max_{k \in \mathcal{Y}} \hat{R}_k(\delta)$ sur $\mathcal{T} = \phi(\mathcal{X})$.

Objectifs et contributions. Bien que le classifieur minimax déterministe $\delta_{\bar{\pi}}^B$ obtienne des résultats très intéressants concernant les objectifs introduits précédemment, il se peut que ce classifieur ne soit pas toujours optimal pour égaliser et minimiser les risques d'erreurs par classe sur un espace partitionné $\mathcal{T} = \phi(\mathcal{X})$. Cette non-optimalité est par exemple illustrée dans la Figure 1. Elle provient de la non-différentiabilité de la surface de Bayes V_B au point $\bar{\pi} = \operatorname{argmax}_{\pi \in \mathbb{S}} V_B(\pi)$ et du fait que le classifieur $\delta_{\bar{\pi}}^B$ soit déterministe. Comme nous pouvons le voir sur la Figure 1, tout classifieur (même non-Bayésien) obtenant des risques conditionnels (2) et globaux (3) bornés dans la surface rose serait plus performant. Dans ce présent article notre contribution est de construire une approche permettant de calculer un tel classifieur $\delta^\star \in \Delta$ vérifiant $\max_{k \in \mathcal{Y}} \hat{R}_k(\delta^\star) \leq \max_{k \in \mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B)$. Puisque le classifieur minimax déterministe $\delta_{\bar{\pi}}^B$ introduit dans [3] est théoriquement démontré comme le classifieur déterministe pour lequel $\max_{k \in \mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B) = \min_{\delta \in \mathcal{D}} \max_{k \in \mathcal{Y}} \hat{R}_k(\delta)$ sur $\mathcal{T} = \phi(\mathcal{X})$, nous devons ainsi nous orienter vers l'ensemble des classifieurs non-déterministes $\Delta \backslash \mathcal{D}$, c'est à dire l'ensemble des classifieurs randomisés. Il est à noter que les classifieurs randomisés ont déjà été étudiés dans le siècle dernier par [1, 2, 7] pour obtenir un classifieur minimax parfaitement égalisateur.

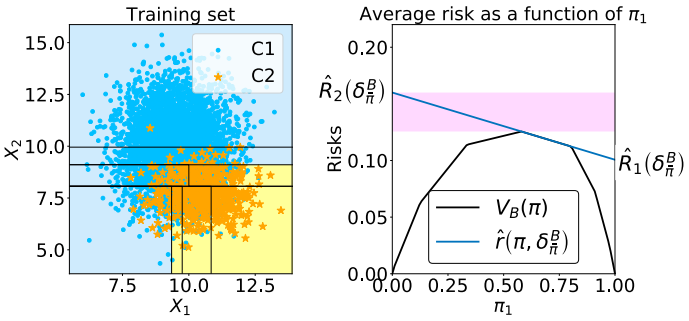

FIGURE 1 : Expériences sur données simulées avec $K = 2$ classes et $d = 2$ variables descriptives numériques ($\mathcal{X} \subset \mathbb{R}^2$). **Gauche.** Frontière de décision du classifieur minimax discret [3] appliqué sur un partitionnement d'arbre de décision $\phi : \mathcal{X} \to \mathcal{T}$ composé de $T = 8$ profils discrets (le nombre de feuilles). **Droite.** La surface V_B correspond au risque de Bayes empirique (8) sur le partitionnement ϕ en fonction des probabilités a priori $\pi \in \mathbb{S}$. Puis que $K = 2$ classes, le risque $\hat{r}(\pi, \delta_{\bar{\pi}}^B)$ défini en (3) peut se réécrire en fonction de π_1 : $\hat{r}(\pi, \delta_{\bar{\pi}}^B) = \pi_1[\hat{R}_1(\delta_{\bar{\pi}}^B) - \hat{R}_2(\delta_{\bar{\pi}}^B)] + \hat{R}_2(\delta_{\bar{\pi}}^B)$.

2 Classifieur Minimax Randomisé

Nous souhaitons approximer le classifieur de Bayes déterministe (6) par un classifieur randomisé $\delta^\star \in \Delta$ qui attribue une classe $k \in \mathcal{Y}$ à partir d'un score de probabilité $\mathbb{P}(\delta^\star(X_i) = k)$. Dans le but de bien approximer le classifieur de Bayes déterministe (6) quelque soit $\pi \in \mathbb{S}$, nous souhaitons ainsi attribuer la classe k à une observation X_i avec une grande probabilité si $k = \operatorname{argmin}_{l \in \mathcal{Y}} f_l(\pi, X_i)$. Pour satisfaire cet objectif, nous proposons de considérer la fonction softmin comme suit.

Définition 1. *En considérant un paramètre de température $\lambda > 0$ ainsi que les quantités $f_1(\pi, X_i), \ldots, f_K(\pi, X_i)$, définies en (7), le classifieur randomisé Softmin-Discret, noté δ_π^λ et associé aux probabilités a priori $\pi \in \mathbb{S}$, attribuera la classe $k \in \mathcal{Y}$ avec probabilité estimée*

$$\hat{\mathbb{P}}\left(\delta_\pi^\lambda(X_i) = k\right) = \frac{e^{-\lambda f_k(\pi, X_i)}}{\sum_{l=1}^{K} e^{-\lambda f_l(\pi, X_i)}}. \tag{9}$$

Ce classifieur randomisé δ_π^λ dépend donc du paramètre de température $\lambda > 0$ ainsi que des probabilités a priori $\pi \in \mathbb{S}$.

La proposition suivante modélise analytiquement le risque empirique du classifieur δ_π^λ sur le simplexe $\mathbb{S}$.

Proposition 1. *En considérant un paramètre de température $\lambda > 0$ fixe, le risque d'erreur global (1) du classifieur Softmin-Discret δ_π^λ et associé aux probabilités a priori $\pi \in \mathbb{S}$ est donné par $V_\lambda : \mathbb{S} \to [0, +\infty)$ telle que*

$$V_\lambda(\pi) := \hat{r}\left(\delta_\pi^\lambda\right) = \sum_{k\in\mathcal{Y}} \pi_k \hat{R}_k\left(\delta_\pi^\lambda\right), \tag{10}$$

où pour chaque classe $k \in \mathcal{Y}$ le risque conditionnel $\hat{R}_k\left(\delta_\pi^\lambda\right)$ est analytiquement donné par

$$\hat{R}_k\left(\delta_\pi^\lambda\right) = \sum_{t\in\mathcal{T}} \sum_{l\in\mathcal{Y}} L_{kl}\, \hat{p}_{kt}\, \frac{e^{-\lambda \sum_{j\in\mathcal{Y}} L_{jl}\, \pi_j\, \hat{p}_{jt}}}{\sum_{q=1}^{K} e^{-\lambda \sum_{j\in\mathcal{Y}} L_{jq}\, \pi_j\, \hat{p}_{jt}}}. \tag{11}$$

Tout comme le classifieur Softmin-Discret δ_π^λ, son risque d'erreur $V_\lambda(\pi)$ dépend aussi du paramètre de température λ. Puisque nous souhaitons que notre classifieur obtienne des risques conditionnels par classe $\hat{R}_k(\delta_\pi^\lambda)$, ainsi que des risques de test moyens $\hat{r}(\cdot, \delta_\pi^\lambda)$, inférieurs à $\max_{k\in\mathcal{Y}} \hat{R}_k\left(\delta_{\bar{\pi}}^B\right)$ (par exemple compris dans la bande rose sur la Figure 1), le paramètre de température λ va jouer un rôle important. La proposition suivante étudie le comportement de δ_π^λ ainsi que de son risque $V_\lambda(\pi)$ en fonction du paramètre de température λ.

Proposition 2. *Le classifieur Softmin-Discret δ_π^λ associé aux probabilités a priori $\pi \in \Pi$ (où $\Pi \subset \mathbb{S}$ correspond aux probabilités a priori pour lesquelles $\operatorname{argmin}_{l\in\mathcal{Y}} f_l(\pi, X_i)$ est unique), converge en probabilité vers le classifieur de Bayes déterministe $\delta_{\bar{\pi}}^B$ (6) lorsque λ tend vers l'infini. De plus, quelque soit $\pi \in \mathbb{S}$, le risque d'erreur moyen $V_\lambda(\pi)$ converge simplement vers le risque de Bayes $V_B(\pi)$ lorsque λ tend vers l'infini.*

Une illustration de la Proposition 2 est proposée dans la Figure 2, Gauche. Alors que le classifieur de Bayes déterministe (6) attribue une unique classe $k \in \mathcal{Y}$ à chaque profil $t \in \mathcal{T}$, le classifieur Softmin-Discret δ_π^λ devient plus souple en se basant sur la règle décision randomisée (9). D'après la Proposition 2, plus λ augmente, plus δ_π^λ devient ferme.

Pour la suite, nous considérerons que le classifieur minimax déterministe $\delta_{\bar{\pi}}^B$ associé aux probabilités a priori $\bar{\pi} = \operatorname{argmax}_{\pi\in\mathbb{S}} V_B(\pi)$ ne permet pas d'obtenir des risques d'erreur par classe égalisés. Ainsi, puisque notre objectif est de construire notre classifieur égalisateur permettant d'obtenir un risque d'erreur moyen entre $V_B(\bar{\pi})$ and $\max_{k\in\mathcal{Y}} \hat{R}_k\left(\delta_{\bar{\pi}}^B\right)$, comme par exemple dans la bande rose de la Figure 1, droite, nous définissons $\Lambda(\bar{\pi})$ l'ensemble des paramètres de température satisfaisant cet objectif :

$$\Lambda(\bar{\pi}) = \left\{ \lambda > 0 : V_B(\bar{\pi}) \leq \max_{\pi\in\mathbb{S}} V_\lambda(\pi) \leq \max_{k\in\mathcal{Y}} \hat{R}_k\left(\delta_{\bar{\pi}}^B\right) \right\}. \tag{12}$$

De plus, pour chaque paramètre de température acceptable $\lambda \in \Lambda(\bar{\pi})$, nous définissons également $\mathcal{B}_\lambda(\bar{\pi})$ l'ensemble des probabilités a priori associées satisfaisant notre objectif :

$$\mathcal{B}_\lambda(\bar{\pi}) = \left\{ \pi \in \mathbb{S} : V_B(\bar{\pi}) \leq V_\lambda(\pi) \leq \max_{k\in\mathcal{Y}} \hat{R}_k\left(\delta_{\bar{\pi}}^B\right) \right\}. \tag{13}$$

Corollaire 1. *Lorsque le classifieur minimax déterministe $\delta_{\bar{\pi}}^B$ associé aux probabilités a priori $\bar{\pi} = \operatorname{argmax}_{\pi\in\mathbb{S}} V_B(\pi)$ n'obtient pas une égalisation des risques d'erreur par classe, les ensembles $\Lambda(\bar{\pi})$ and $\mathcal{B}_\lambda(\bar{\pi})$, $\lambda \in \Lambda(\bar{\pi})$, sont non-vides.*

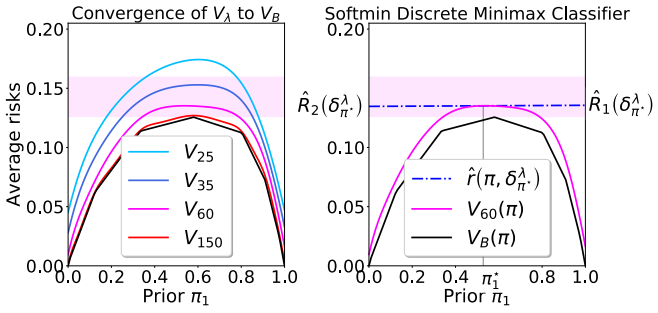

FIGURE 2 : Expériences sur la base de données synthétique introduite dans la Figure 1. La bande rose correspond à l'ensemble des risques moyens (1), (3), bornés par $V_B(\bar{\pi})$ et $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B)$. **Gauche.** Lorsque λ augmente, $V_\lambda(\pi)$ converge simplement vers le risque de Bayes $V_B(\pi)$. **Droite.** Pour $\lambda = 60 \in \Lambda(\bar{\pi})$, $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ permet au classifieur randomisé $\delta_{\pi^\star}^\lambda$ d'égaliser les risques conditionnels et d'obtenir des risques d'erreur moyens (1), (3) inférieurs à $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B)$.

Nous pouvons observer sur la Figure 2, gauche, que pour chaque paramètre de température acceptable $\lambda \in \Lambda(\bar{\pi})$ et quelque soit $\pi \in \mathcal{B}_\lambda(\bar{\pi})$, $V_\lambda(\pi)$ appartient à la bande rose, ce qui est l'un de nos objectifs à atteindre. Il nous reste maintenant à démontrer que quelque soit $\lambda \in \Lambda(\bar{\pi})$, il existe $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ tel que le classifieur randomisé Softmin-Discret $\delta_{\pi^\star}^\lambda$ permet d'égaliser tous ses risques d'erreur par classe, comme par exemple dans la Figure 2, droite.

Pour cela, considérons l'application $G : \mathbb{S} \to \mathbb{R}^K$ définie par $G(\pi) := [g_1(\pi), \ldots, g_K(\pi)]$ où quelque soit $k \in \mathcal{Y}$,

$$g_k(\pi) := \hat{R}_k\left(\delta_\pi^\lambda\right) - V_\lambda(\pi). \tag{14}$$

L'application $G(\pi)$, qui est analytiquement calculable à partir des équations (10) et (11), mesure l'écart entre le risque d'erreur moyen $V_\lambda(\pi)$ et chaque risque d'erreur conditionnel par classe $\hat{R}_k\left(\delta_\pi^\lambda\right)$. Le lemme suivant fourni une condition nécessaire et suffisante établissant que le classifieur Softmin-Discret $\delta_{\pi^\star}^\lambda$ permet d'égaliser tous ses risques d'erreur par classe $\hat{R}_k\left(\delta_\pi^\lambda\right)$.

Lemme 1. *Pour chaque paramètre de température $\lambda \in \Lambda(\bar{\pi})$, le classifieur randomisé Softmin-Discret $\delta_{\pi^\star}^\lambda$ associé aux probabilités a priori $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ égalise tous ses risques d'erreurs conditionnels par classe si et seulement si $G(\pi^\star) = 0$. De plus, quelque soit $\lambda \in \Lambda(\bar{\pi})$, une telle racine $\pi^\star$ existe dans l'ensemble $\mathcal{B}_\lambda(\bar{\pi})$.*

Alors que le classifieur minimax déterministe $\delta_{\bar{\pi}}^B \in \mathcal{D}$ cherche à minimiser le maximum des risques conditionnels par classe sur un espace des variables descriptives partitionné, le théorème suivant établit que notre classifieur minimax randomisé Softmin-Discret atteint de meilleures performances pour cet objectif sur ce même espace partitionné.

Théorème 1. *Si les risques par classe $\hat{R}_k(\delta_{\bar{\pi}}^B)$ du classifieur minimax déterministe $\delta_{\bar{\pi}}^B \in \mathcal{D}$ ne sont pas tous égalisés sur l'espace des variables descriptives partitionné $\mathcal{T} = \phi(\mathcal{X})$, alors quelque soit $\lambda \in \Lambda(\bar{\pi})$, le classifieur minimax randomisé Softmin-Discret $\delta_{\pi^\star}^\lambda$ associé aux probabilités a priori $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ satisfaisant $G(\pi^\star) = 0$, permet d'obtenir, sur ce même espace partitionné $\mathcal{T}$: $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta_{\pi^\star}^\lambda) \leq \max_{k\in\mathcal{Y}} \hat{R}_k(\delta_{\bar{\pi}}^B)$.*

D'après le Lemme 1 et le Théorème 1, pour un paramètre de température $\lambda \in \Lambda(\bar{\pi})$ fixé, il nous suffit donc de calculer une racine $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ satisfaisant $G(\pi^\star) = 0$ de sorte que le classifieur minimax randomisé $\delta^\lambda_{\pi^\star}$ égalise tous ses risques conditionnels par classe et que $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta^\lambda_{\pi^\star}) \leq \max_{k\in\mathcal{Y}} \hat{R}_k(\delta^B_{\bar{\pi}})$, comme illustré sur la Figure 2, droite. Afin de calculer cette racine $\pi^\star \in \mathcal{B}_\lambda(\bar{\pi})$ satisfaisant $G(\pi^\star) = 0$, nous proposons de considérer le problème d'optimisation

$$\pi^\star = \operatorname*{argmin}_{\pi\in\mathbb{S}} \|G(\pi)\|_2^2. \tag{15}$$

Ce problème de minimisation n'étant pas convexe, nous pouvons le résoudre en considérant un algorithme à base de descente de gradient stochastique comme proposé par [5], ou par une méthode de Monte-Carlo comme par recuit-simulé [4] en utilisant la distribution de Dirichlet.

Finalement, notre classifieur minimax randomisé Softmin-Discret $\delta^\lambda_{\pi^\star}$ attribuera la classe $l \in \mathcal{Y}$ à partir de la règle de décision (9) en considérant les probabilités a priori $\pi^\star$ solution de (15). Remarquons que nous utilisons ici le terme *"minimax randomisé Softmin-Discret"* pour caractériser le fait que ce classifieur quasi-Bayésien obtient de meilleures performances que le véritable classifieur minimax (Bayesien) déterministe dans la minimisation du maximum des risques par classe.

3 Expériences numériques

Base de données. Considérons la base de données réelles Scania Trucks [8] pour laquelle l'objectif est de prédire la panne d'un composant mécanique de l'APS à partir de 130 mesures numériques. On considère donc $K = 2$ classes, où la classe 1 correspond à l'absence de panne, et la classe 2 caractérise les APS présentant une défaillance. La fonction de perte L est imposée par les experts du domaine d'application telle que $L_{11} = 0$, $L_{12} = 10$, $L_{21} = 500$ et $L_{22} = 0$. Les experts ont fourni un échantillon d'apprentissage et un échantillon de test contenant respectivement 54731 et 14578 observations. Les proportions par classe sont très déséquilibrées : $\hat{\pi} = [0.99, 0.01]$. Ainsi, la détection d'une panne se veut naturellement très difficile.

Procédure expérimentale. Nous avons réalisé une validation croisée et nous avons comparé notre nouvelle approche avec trois autres méthodes adaptées pour équilibrer les risques conditionnels par classe : la Régression Logistique Repondérée (WLR), les arbres de décision repondérés (WDT), et le classifieur Minimax Discret [3] (DMC). Les méthodes WLR et WDT ont été calibrées en utilisant Scikit-Learn [6]. À chaque itération de la validation croisée, nous avons appliqué le DMC et notre nouveau classifieur minimax Softmin-Discret Discret (SoftminDMC) sur le même partitionnement des variables descriptives de chaque sous-ensemble d'apprentissage.

Résultats. Les résultats sont présentés dans la Table 1. Nous pouvons observer en illustration du Théorème 1 que notre classifieur SoftminDMC est plus performant que le DMC pour minimiser le maximum des risques conditionnels par classe et pour équilibrer ces risques par classe. De plus, il permet de diviser le risque de non-identifier une panne par plus de trois comparé aux méthodes usuelles WDT et WLR.

TABLE 1 : Résultats présentés comme $[\text{mean} \pm \text{std}]$. Le critère $\psi(\delta) = \max_{k\in\mathcal{Y}} \hat{R}_k(\delta) - \min_{k\in\mathcal{Y}} \hat{R}_k(\delta)$ mesure la qualité de chaque classifieur δ pour égaliser les risques par classe.

CRITÈRES	CLASSIFIEURS $\delta \in \Delta$	SCANIA-TRUCK Train	SCANIA-TRUCK Test
$\hat{r}(\delta)$	WLR	0.70 ± 0.02	0.84 ± 0.13
	WDT	0.69 ± 0.01	0.86 ± 0.05
	DMC	8.95 ± 0.79	9.00 ± 0.76
	SoftminDMC	4.81 ± 0.42	4.83 ± 0.43
$\max_{k\in\mathcal{Y}} \hat{R}_k(\delta)$	WLR	37.7 ± 2.68	51.2 ± 16.9
	WDT	17.7 ± 5.70	31.9 ± 8.79
	DMC	9.05 ± 0.80	9.16 ± 0.74
	SoftminDMC	5.06 ± 0.49	6.91 ± 1.61
$\psi(\delta)$	WLR	37.4 ± 2.68	50.9 ± 16.9
	WDT	17.2 ± 5.77	31.5 ± 8.86
	DMC	9.05 ± 0.80	5.03 ± 3.66
	SoftminDMC	0.41 ± 0.35	2.83 ± 1.35

4 Conclusion et enjeux sociétaux

Nous avons proposé une nouvelle approche pour détecter des classes rares ou lorsque les probabilités a priori peuvent évoluer au cours du temps. Dans le cas où l'aspect randomisé n'est pas envisageable, notre approche pourrait tout de même être utilisée en règle MAP : elle fournirait ainsi des prédictions proche du classifieur minimax déterministe [3] tout en fournissant des scores de confiance basés sur l'estimation (9).

Références

[1] James O. BERGER : *Statistical decision theory and Bayesian analysis ; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.

[2] T.S. FERGUSON : *Mathematical Statistics : A Decision Theoretic Approach*. Academic Press, 1967.

[3] Cyprien GILET, Susana BARBOSA et Lionel FILLATRE : Discrete box-constrained minimax classifier for uncertain and imbalanced class proportions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] Andrea LECCHINI-VISINTINI, John LYGEROS et Jan MACIEJOWSKI : Simulated annealing : Rigorous finite-time guarantees for optimization on continuous domains. *Advances in Neural Information Processing Systems*, 2007.

[5] Zhize LI, Hongyan BAO, Xiangliang ZHANG et Peter RICHTÁRIK : Page : A simple and optimal probabilistic gradient estimator for nonconvex optimization. *In International Conference on Machine Learning*, 2021.

[6] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT et et AL : Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011.

[7] H. V. POOR : *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 2nd édition, 1994.

[8] CV. AB. SCANIA : APS failure at Scania Trucks data set. https ://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set, 2016.