Décider à partir d'exemples de test partiellement connus

Benjamin Quost

Décider à partir d'exemples de test partiellement connus

Benjamin Quost
UMR UTC-CNRS 7253 Heudiasyc, Université de Technologie de Compiègne ; benjamin.quost@hds.utc.fr

Résumé :
Nous considérons le problème de classer des exemples de test partiellement connus, leurs valeurs manquantes pouvant être identifiées par requêtes successives. Le budget de requêtes est limité : le problème est de se concentrer sur les caractéristiques les plus intéressantes. Nous étudions une stratégie où la distribution des caractéristiques restantes est mise à jour au fur et à mesure des valeurs découvertes. Après quelques expériences, nous discutons du cas de réponses imprécises, et mentionnons les problèmes qui se posent alors.

Mots-clés :
Prise de décision partiellement supervisée ; prise de décision robuste ; apprentissage actif.

Abstract:
We consider classifying partially uncovered test instances, by uncovering (some of) their missing values via successive requests. The request budget is limited : the problem is then to concentrate on the most informative features. We study a strategy where the feature distribution is updated according to the uncovered values. After a few experiments, we briefly discuss the case of imprecise answers, and point out the difficulties which then arise.

Keywords:
Partially supervised decision-making ; robust decision-making ; active learning.

1 Introduction

Dans un problème classique de classification, un classifieur est entraîné à identifier la classe d'exemples tirés d'une population de référence. Ces exemples sont interprétés comme les réalisations d'un vecteur aléatoire $\boldsymbol{X} \in \mathbb{R}^p$ dont les éléments sont les caractéristiques descriptives des exemples. La théorie bayésienne de la décision [4, 13] justifie de classer ces exemples à partir des probabilités a posteriori des classes. Les approches discriminatives estiment ces probabilités directement, par exemple via un modèle paramétré appris à partir d'exemples d'apprentissage ; les approches génératives [12] estiment les distributions conditionnelles des classes et leurs probabilités a priori, puis déduisent les probabilités a posteriori en utilisant la règle de Bayes.

Dans les deux cas, la prise de décision requiert la connaissance parfaite de l'exemple à classer. Dans certaines applications, cependant, les exemples peuvent être partiellement observés. Par exemple, en diagnostic médical, rien n'est connu de l'état d'un patient avant qu'il ne soit examiné, et identifier certaines caractéristiques peut requérir de faire des examens invasifs et potentiellement dangereux.

Nous supposons dans cet article que les exemples de test ne sont pas observés : leurs caractéristiques manquantes doivent être découvertes pour classer l'exemple. Ce problème diffère de l'imputation en ce que les valeurs manquantes sont découvertes et non estimées, et se rapproche en cela de l'apprentissage actif [5, 10, 3, 15, 14], qui considère cependant la mise à jour d'un modèle en phase d'apprentissage, et vise généralement à identifier la variable de classe pour certains exemples par ailleurs connus. Il est à noter que dans [15, 14], le problème de l'identification de caractéristiques manquantes est considéré, et se rapproche de notre proposition dans l'esprit. Dans [9], le problème de classer des exemples de test partiels est considéré. Cependant, les caractéristiques manquantes sont supposées être toujours les mêmes, et l'objectif est de choisir de compléter un exemple pour le classer avec un modèle complet, ou de le laisser partiel et de le classer avec un modèle ne nécessitant que les caractéristiques toujours disponibles.

La stratégie proposée ici consiste à découvrir les valeurs manquantes de manière itérative, dans le but d'accumuler les éléments pointant vers l'une des classes en présence, en utilisant le modèle appris préalablement. Un oracle est supposé fournir les valeurs demandées, de manière infaillible. Le cas de réponses imprécises ou in-

certaines, et les conséquences sur la stratégie proposée ici, seront brièvement évoqués.

Le paragraphe 2 rappelle brièvement le problème et introduit les notations. Le paragraphe 3 décrit l'approche proposée, le paragraphe 4 présente des résultats d'expériences préliminaires, et le paragraphe 5 discute brièvement du cas de réponses imprécises. Le paragraphe 6 conclut en présentant quelques directions de recherche futures.

2 Cadre

2.1 Prise de décision classique

On cherche à affecter un exemple à une classe parmi $K \geq 2$ classes en présence. L'exemple est supposé être la réalisation d'un vecteur aléatoire $\boldsymbol{X} \in \mathcal{X} = \mathbb{R}^p$, et l'information de classe est encodée par une variable $Z \in \Omega = \{\omega_1, \ldots, \omega_K\}$. Nous considérons ici un modèle génératif : les probabilités a priori $\pi_k = \Pr(Z = \omega_k)$ et les densités conditionnelles $f_k = f_{\boldsymbol{X}|Z=\omega_k}$ ont été estimées à partir de données d'apprentissage. Elles peuvent être utilisées pour calculer les probabilités a posteriori des classes pour l'exemple à classer :

$$\omega^\star = \arg \max_{k=1,\ldots,K} \Pr(\omega_k|\boldsymbol{x}), \qquad (1)$$

où

$$\Pr(\omega_k|\boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_\ell \pi_\ell f_\ell(\boldsymbol{x})}. \qquad (2)$$

2.2 Décider à partir d'exemples incomplets

Supposons que les exemples de test soient partiellement observés. Pour chacun, on peut décomposer le vecteur de caractéristiques en deux : $\boldsymbol{x} = (\boldsymbol{x}_O, \boldsymbol{x}_M)$, où $O, M \subseteq \{1, \ldots, p\}$ sont respectivement les ensembles des indices des variables observées et manquantes (ils forment une partition de $\{1, \ldots, p\}$). Dès lors que $M \neq \emptyset$, les densités conditionnelles $f_k(\boldsymbol{x})$ ne peuvent être calculées au point $\boldsymbol{x}$; une décision peut être prise à partir de $\boldsymbol{x}_O$ seul, en marginalisant les variables manquantes. Ces informations pouvant être insuffisantes pour prendre de bonnes décisions, on peut chercher à identifier une partie des données manquantes dans $\boldsymbol{x}_M$ avant de procéder au classement.

On propose ici d'identifier les caractéristiques manquantes une à une par requêtes successives, en exploitant l'information récupérée pour identifier les prochaines requêtes par mise à jour des densités conditionnelles au point $\boldsymbol{x}_O$.

3 Processus itératif de requêtage

3.1 Stratégie globale

On propose de découvrir les caractéristiques manquantes du vecteur à classer une à une. À chaque étape $t = 0, 1, 2, \ldots$ du processus, on dispose d'ensembles de caractéristiques observées O_t et manquantes M_t. Le processus, décrit par l'algorithme 1, consiste à identifier des variables Q_t de M_t et les transférer à O_{t+1}.

Algorithme 1 : Requêtage itératif

Input : exemple $\boldsymbol{x}$ avec valeurs observées O_t et manquantes M_t ; nombre $q_{\max}$ de requêtes ; modèle génératif

Output : Ensembles O_t et M_t mis à jour

for *t=1,2,...,*$q_{\max}$ **do**

- identifier un nouvel élément $\boldsymbol{x}_{Q_t}$ de $\boldsymbol{x}$ à découvrir;
- mettre à jour les ensembles : $O_t \leftarrow O_{t-1} \cup Q_t$, $M_t \leftarrow M_{t-1} \setminus Q_t$;
- mettre à jour la distribution des M_t caractéristiques restantes;

return ensembles O_t et M_t, vecteur de caractéristiques observées $\boldsymbol{x}_{O_t}$

Initialement, aucune valeur n'est connue ($O_0 = \emptyset$, $M_0 = \{1, \ldots, p\}$) : l'exemple devrait être classé dans la classe de probabilité a priori maximale. Plutôt que d'identifier toutes les valeurs manquantes, on peut n'en n'identifier qu'un sous-ensemble Q_1 : $O_1 \leftarrow Q_1$ et $M_1 \leftarrow \{1, \ldots, p\} \setminus Q_1$. On peut alors prendre une décision, ou formuler une nouvelle requête si le budget le permet.

Le principal problème est d'identifier les caractéristiques à propos desquelles interroger l'expert : lesquelles, combien à la fois, dans quel ordre. Nous proposons de quantifier la quantité d'information susceptible d'être apportée par chaque caractéristique. Sur la base de ce critère d'informativité, nous proposons d'identifier les caractéristiques les plus informatives, en répétant la procédure de manière itérative si le budget de questions le permet.

3.2 Critère d'informativité

Intuitivement, le choix des variables à identifier devrait être fait en fonction de leur propension à séparer les classes. Une variable est informative si elle a un rôle important dans la prédiction de la variable de sortie Z. En conséquence, ayant observé précédemment les valeurs dans $\boldsymbol{x}_O$, nous proposons d'interroger l'oracle à propos d'une nouvelle variable $\boldsymbol{X}_q$ qui minimise *l'entropie conditionnelle* $H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O)$ [11] :

$$H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O) = \\ -\sum_{k=1}^{K} \int_{\mathcal{X}_q} \Pr(\omega_k, \boldsymbol{x}_q|\boldsymbol{x}_O) \log \frac{\Pr(\omega_k, \boldsymbol{x}_q|\boldsymbol{x}_O)}{f_{\boldsymbol{X}_q|\boldsymbol{x}_O}(\boldsymbol{x}_q)} d\boldsymbol{x}_q. \tag{3}$$

Notons que cette entropie conditionnelle, qui indique dans quelle mesure $\boldsymbol{X}_q$ détermine la variable de classe Z, peut s'écrire comme l'entropie différentielle espérée de $Z|\boldsymbol{x}_q, \boldsymbol{x}_O$, qui est l'entropie de la distribution $\Pr(Z|\boldsymbol{x}_q, \boldsymbol{x}_O)$:

$$H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O) = \mathbb{E}_{\boldsymbol{X}_q|\boldsymbol{x}_O}\left[H(Z|\boldsymbol{x}_q, \boldsymbol{x}_O)\right]. \tag{4}$$

Cette stratégie revient donc à choisir la variable telle que la masse de probabilité a posteriori sera distribuée le plus possible vers une classe unique. Remarquons enfin que minimiser l'entropie revient à maximiser une divergence de Kullback-Leibler espérée entre les densités conditionelles et la densité de mélange :

$$H(Z|\boldsymbol{X}_q, \boldsymbol{X}_O = \boldsymbol{x}_O) = H(Z) \\ \underbrace{-\sum_k \pi_k \int_{\mathcal{X}_q} f_k(\boldsymbol{x}_q|\boldsymbol{x}_O) \log \frac{f_k(\boldsymbol{x}_q|\boldsymbol{x}_O)}{f(\boldsymbol{x}_q|\boldsymbol{x}_O)} d\boldsymbol{x}_q}_{\mathbb{E}_Z[\mathrm{D_{KL}}(\Pr(\boldsymbol{X}_q|Z,\boldsymbol{x}_O)||\Pr(\boldsymbol{X}_q|\boldsymbol{x}_O))]}.$$

On choisit donc la variable qui maximise le gain d'information induit par l'utilisation de la distribution marginale $\Pr(\boldsymbol{X}_q|\boldsymbol{x}_O)$ à la place de la distribution conditionnelle $\Pr(\boldsymbol{X}_q|Z, \boldsymbol{x}_O)$ (moyennée par rapport à la distribution de Z).

3.3 Calcul pratique de l'entropie

Le calcul pratique de l'entropie requiert un modèle de la distribution jointe $\Pr(\boldsymbol{X}_q, Z|\boldsymbol{x}_O)$ pour tout $q \in \{1, \ldots, p\}$. Dans le cas d'un modèle de classification génératif, on peut obtenir une (estimation de) cette distribution jointe, comme par exemple dans le cas gaussien. Dans le cas de modèles de classification discriminatifs, cela ne sera pas le cas.

Cependant, même avec un modèle génératif, $H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O)$ ne peut en général être calculée exactement. Remarquons que (3) peut être réécrite comme

$$H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O) = H(\boldsymbol{X}_q|Z, \boldsymbol{x}_O) \\ + H(Z) - H(\boldsymbol{X}_q|\boldsymbol{x}_O); \tag{5}$$

si $H(\boldsymbol{X}_q|Z, \boldsymbol{x}_O)$ peut admettre une expression formelle — cela dépend de la distribution conditionnelle $f_k(\cdot|\boldsymbol{x}_O)$, ça ne sera généralement pas le cas de $H(\boldsymbol{X}_q|\boldsymbol{x}_O)$, $f_{\boldsymbol{X}_q}(\cdot|\boldsymbol{x}_O)$ étant un mélange de distributions.

On pourra toujours calculer une approximation de (4), ou des termes de (5) pour lesquels il n'existe pas d'expression formelle, via des stratégies de Monte-Carlo ; par exemple :

$$H(Z|\boldsymbol{X}_q, \boldsymbol{x}_O) \simeq \frac{1}{T} \sum_{t=1}^{T} H(Z|\boldsymbol{x}_O, \boldsymbol{x}_q^{(t)}),$$

où les T exemples $\boldsymbol{x}_q^{(1)}, \ldots, \boldsymbol{x}_q^{(T)}$ sont générés suivant $f_{\boldsymbol{X}_q|\boldsymbol{x}_O}(\cdot) = \sum_k \pi_k f_k(\cdot|\boldsymbol{x}_O)$. Empiriquement, les estimations semblent plus précises lorsqu'on approche (4) plutôt que (5).

3.4 Mise à jour des distributions

Dans l'algorithme 1, la dernière étape d'une itération consiste à mettre à jour la distribution des caractéristiques restant à identifier en

fonction des valeurs qui viennent d'être observées. Si la réponse d'un oracle est une valeur précise $\boldsymbol{x}_q$ pour la variable $\boldsymbol{X}_q$, une stratégie triviale et naïve consisterait en effet à simplement transférer la valeur obtenue de M_t à O_t, et d'identifier la prochaine requête indépendamment de cette valeur $\boldsymbol{x}_q$. Cette procédure est naïve en ce qu'elle revient à calculer les entropies conditionnelles des variables $\boldsymbol{X}_q$ séparément avant d'effectuer les différentes itérations de l'algorithme, et néglige donc les interactions entre variables.

En effet, tout élément d'information $\boldsymbol{x}_q$ communiqué par l'expert est à même de modifier notre connaissance de la distribution des variables restant à identifier, en particulier si les variables sont fortement corrélées : il est possible de mettre à jour ces distributions en les conditionnant par rapport à $\boldsymbol{x}_q$. Les requêtes ultérieures seront donc impactées par le résultat de la requête courante $\boldsymbol{x}_q$. Cette étape de mise à jour peut être délicate, selon les distributions conditionnelles considérées. La famille gaussienne (multivariée), utilisée ici, est stable par conditionnement, les seuls changements intervenant au niveau des paramètres (pour lesquels il existe des formules de mise à jour).

3.5 Complexité calculatoire

Le coût de la stratégie naïve est faible : l'entropie n'est estimée pour chaque variable qu'une seule fois, avant la phase de test. La stratégie séquentielle, elle, nécessite de ré-estimer les entropies après chaque mise à jour des distributions conditionnelles : pour chaque X_q, il faut

1. générer un échantillon de valeurs $x_q^{(\ell)}$ (avec $\ell = 1, \ldots, L$, où L est choisi),
2. recalculer les probabilités a posteriori des classes aux points $(\boldsymbol{x}_{O_t}, x_q^{(\ell)})$,
3. estimer l'entropie en moyennant les entropies de ces probabilités a posteriori.

Le coût dépend directement du nombre L de valeurs générées pour estimer l'entropie espérée. Dans le cas où le modèle génératif ne donnerait qu'une approximation grossière des vraies distributions conditionnelles, l'approche naïve, plus robuste, peut s'avérer un meilleur choix que l'approche séquentielle.

4 Expériences

4.1 Exemple simple

Nous illustrons la stratégie décrite ci-dessus sur un exemple simple. On considère une population d'exemples $\boldsymbol{x}_i \in \mathbb{R}^3$ répartis en $K = 2$ classes gaussiennes, présentes en proportions $\pi_1 = 0.5 = \pi_2$, les paramètres étant

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1.5 \\ -0.5 \\ -0.5 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} -1.5 \\ 0.5 \\ 0.5 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.75 & 0.25 \\ 0.75 & 1 & 0 \\ 0.25 & 0 & 1 \end{pmatrix}.$$

L'entropie est estimée par la procédure décrite au paragraphe 3.3, équation (4).

La première requête porte sur la variable X_1. Intuitivement, c'est la plus discriminante (comme le confirme la figure 1), ce que confirment les entropies estimées : $H(Z|X_1) \simeq 0.166$, tandis que $H(Z|X_2) = H(Z|X_3) \simeq 0.581$.

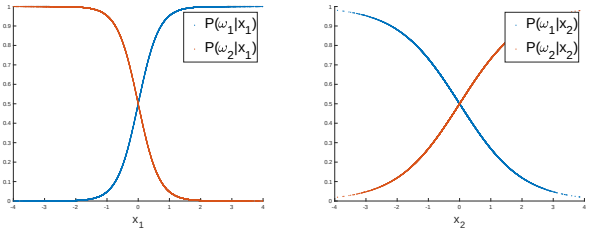

Figure 1 – Probabilités a posteriori $\Pr(\omega_k|x_j)$ pour $j = 1$ (gauche), et $j \in \{2, 3\}$ (droite).

Supposons que la valeur observée pour X_1 soit $x_1 = 0.5$. S'il fallait décider, les probabilités a posteriori amèneraient à choisir ω_1 : en effet, $\Pr(\omega_1|X_1 = 0.5) \simeq 0.818$. D'autres valeurs de variables manquantes peuvent néanmoins être identifiées. D'après les entropies précédemment

calculées, X_2 et X_3 sont a priori aussi informatives l'une que l'autre. Mettons à jour les distributions de ces variables : notons $\boldsymbol{X}_{23|1}$ le vecteur aléatoire $(X_2, X_3)^T | X_1 = x_1$. On a

$$\boldsymbol{X}_{23|1} \underset{\omega_k}{\sim} \mathcal{N}\left(\boldsymbol{\mu}_{k,23|1}, \boldsymbol{\Sigma}_{k,23|1}\right),$$

avec

$$\boldsymbol{\mu}_{1,23|1} = \begin{pmatrix} -1.25 \\ -0.75 \end{pmatrix}, \quad \boldsymbol{\mu}_{2,23|1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{1,23|1} = \boldsymbol{\Sigma}_{2,23|1} = \begin{pmatrix} 0.4375 & -0.1875 \\ -0.1875 & 0.9375 \end{pmatrix}.$$

La variable $X_2|x_1$ est à présent la plus discriminante (voir figure 2) : $H(Z|X_2, x_1) \simeq 0.0225$, et $H(Z|X_3, x_1) \simeq 0.3637$.

Supposons que nous observons $X_2 = x_2 = 2$, ce qui amènerait à affecter l'exemple à la classe ω_2, avec $\Pr(\omega_1|X_1 = 0.5, X_2 = 2) \simeq 0$. Si $X_3 = x_3 = 0.5$ avait été observé, la classe ω_1 aurait été choisie avec $\Pr(\omega_1|X_1 = 0.5, X_3 = 0.5) \simeq 0.69$. La décision optimale, basée sur le vecteur complet, est ω_2 ($\Pr(\omega_1|\boldsymbol{x}) \simeq 0$).

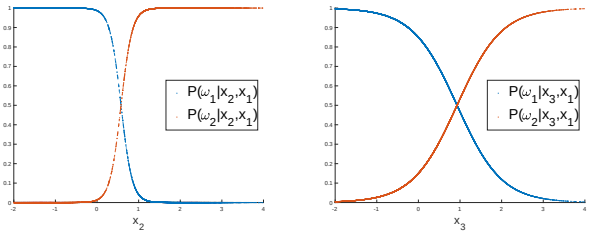

Figure 2 – Probabilités $\Pr(\omega_k|x_j, x_1)$ calculées pour $j = 2$ (gauche) et $j = 3$ (droite).

4.2 Données synthétiques

Nous avons généré $n = 1000$ exemples selon un mélange de gaussiennes avec $\pi_1 = 0.6$, et

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1, 1, 1, 1, 1 \end{pmatrix}^T, \quad \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1;$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.25 & 0 & 0 & 0 \\ 0.25 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.75 & 0.25 \\ 0 & 0 & 0.75 & 1 & 0.5 \\ 0 & 0 & 0.25 & 0.5 & 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.25 & 0.5 & 0 & 0 \\ 0.25 & 1 & 0.75 & 0 & 0 \\ 0.5 & 0.75 & 1 & 0.25 & 0 \\ 0 & 0 & 0.25 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}.$$

Les données ont été réparties aléatoirement en ensembles d'apprentissage (60%), sur lequel nous avons estimé un modèle d'analyse discriminante quadratique, et de test (40%). Nous avons comparé quatre approches de classement, selon les caractéristiques utilisées :

1. toutes sont exploitées ;
2. un sous-ensemble aléatoire de n_f variables est utilisé pour chaque exemple ;
3. approche naïve : n_f variables sont choisies sur la base des entropies $H(Z|X_j)$ estimées sans mise à jour ;
4. approche séquentielle : n_f variables sont choisies grâce aux entropies mises à jour et fur au à mesure des requêtes.

Pour les approches 2, 3 et 4, nous avons considéré $n_f = 1, \ldots, p-1$. L'entropie est estimée par Monte-Carlo (4) avec 10^4 exemples. Nous avons répété cette procédure $T = 25$ fois.

La figure 3 montre l'évolution du taux d'erreur moyen (avec un intervalle de confiance à 95%). Lorsque $n_f = 1$, les stratégies 3 and 4 donnent les mêmes résultats (les différences sont dues à l'approximation de l'entropie). La stratégie séquentielle est significativement plus efficace que la sélection aléatoire, et meilleure que l'approche naïve (la différence étant significative pour $n_f = 2$ et $n_f = 3$). Les performances se rapprochent évidemment de la stratégie utilisant toutes les variables quand n_f augmente.

Par ailleurs, nous avons constaté que la première variable requêtée était $X_{q_1} = X_3$, pour des raisons numériques (techniquement, toutes les variables ont le même pouvoir discriminant). Nous avons estimé les fréquences de chaque séquence de requêtes faite avec la stratégie séquentielle, selon la vraie classe de l'exemple considéré. Les séquences les plus fréquentes sont données dans les tableaux 1 (classe 1) et 2 (classe 2).

On remarquera que les deux classes ont une requête fréquente en commun — on peut imaginer qu'il s'agit d'exemples à la frontière des deux classes. Les 2e et 3e requêtes semblent ensuite caractéristiques de la classe. Il est intéressant de noter que les deux séquences les plus fréquentes pour la classe 1 *ne sont jamais retrouvées* pour des exemples de la classe 2 (les deux séquences les plus fréquentes pour la classe 2 étant calculées pour 10 et 8 instances de la classe 1). Il pourrait donc exister des séquences "typiques" pour les exemples d'une classe particulière. Une idée pourrait être de précalculer un certain nombre de telles séquences pour chaque classe, pour ensuite choisir une séquence en fonction d'un critère de similarité, afin de limiter le coût calculatoire de la stratégie séquentielle présentée ci-dessus.

Tableau 1 – Séquences les plus fréquentes, exemples de la classe ω_1

Séquence	nb.	fréquence
(3, 1, 2, 4, 5)	33	14.04%
(3, 1, 2, 5, 4)	29	12.34%
(3, 4, 2, 5, 1)	25	10.64%

Tableau 2 – Séquences les plus fréquentes, exemples de la classe ω_2

Séquence	nb.	fréquence
(3, 2, 4, 5, 1)	46	28.40%
(3, 2, 4, 1, 5)	26	16.05%
(3, 4, 2, 5, 1)	22	13.58%

4.3 Données réelles

Nous présentons ici des résultats obtenus sur quatre jeux de données réelles de l'UCI [7]. Pour chacun, nous avons utilisé la procédure décrite au paragraphe 4.2. Les données ont été centrées et réduites, et pour les données *optdigits* et *satimage* nous avons sélectionné $p =$

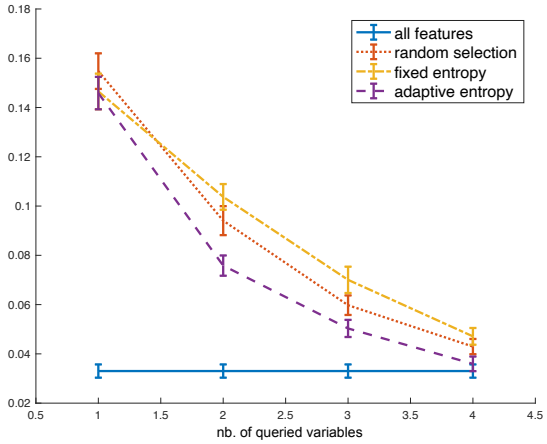

Figure 3 – Courbes d'erreur en fonction du nombre de requêtes, données synthétiques.

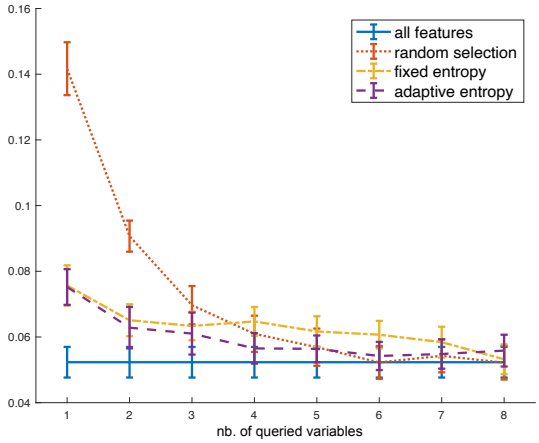

Figure 4 – Courbes d'erreur, *breastcancer*.

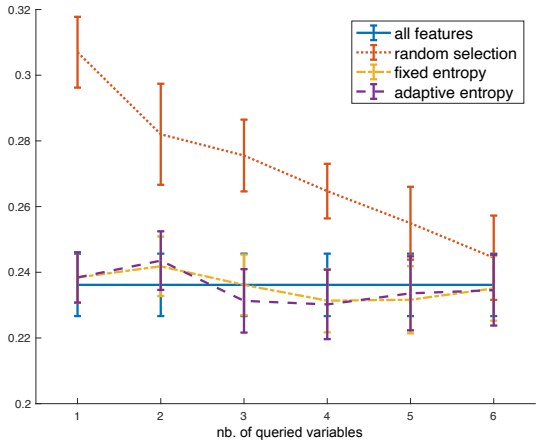

Figure 5 – Courbes d'erreur, données *pima*.

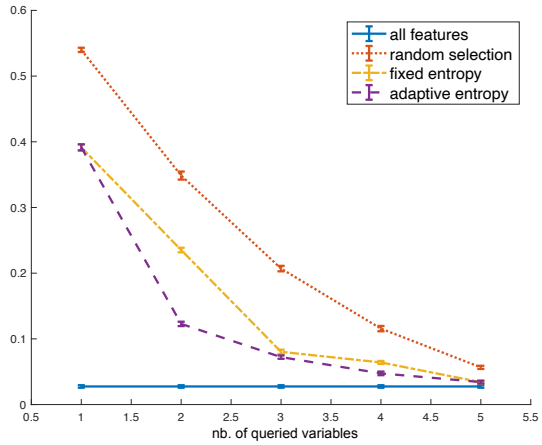

Figure 6 – Courbes d'erreur, données *optdigits*.

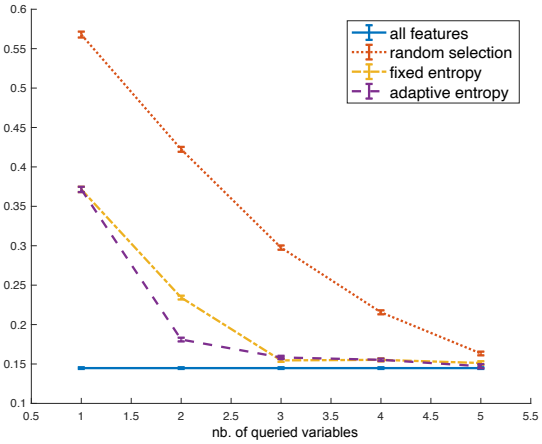

Figure 7 – Courbes d'erreur, données *satimage*.

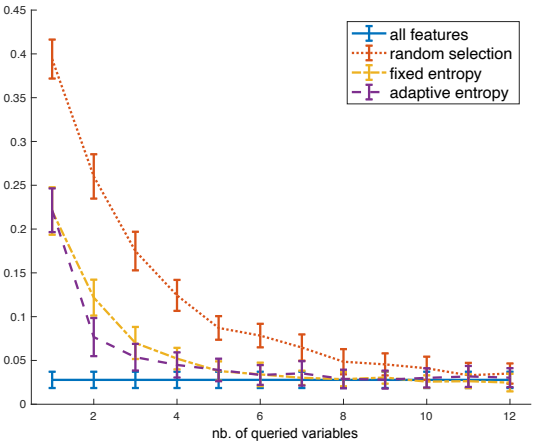

Figure 8 – Courbes d'erreur, données *wine*.

6 variables descriptives par ACP pour éviter les problèmes numériques lors de l'estimation des matrices de covariance de l'ADQ. Pour les données *optdigits*, nous avons considéré cinq classes (chiffres '0', '1', '3', '7', '8') pour limiter le coût calculatoire de l'expérience.

Les figures 4 à 7 montrent l'erreur de classification en fonction du nombre de variables découvertes. Ces courbes confirment que sélectionner les variables en fonction de leur pouvoir discriminant espéré semble fructueux, en particulier si le nombre de requêtes possibles est limité. La procédure séquentielle ne semble toutefois pas systématiquement meilleure que la procédure naïve ; et le cas échéant, la différence est peu souvent significative (*optdigits*, *satimage* et *wine*, avec $n_f = 2$).

Soulignons que si l'hypothèse de normalité des données n'est pas vérifiée, la mise à jour des paramètres dans la procédure séquentielle débouche vraisemblablement sur des distributions très différentes d'une distribution normale (cela impacte également l'estimation de l'entropie). Cela peut mener à formuler des requêtes sous-optimales. Cela suggère que l'usage de méthodes robustes basées sur des ensembles de distributions pourrait être une stratégie opportune dans le cas de données réelles.

5 Réponses imprécises

5.1 Problème

La stratégie séquentielle requiert des réponses précises de la part de l'oracle, sous la forme d'observations $\boldsymbol{X}_q = \boldsymbol{x}_q$. Cela semble réaliste pour certaines applications (ex. diagnostic médical). Pour diverses raisons, les réponses aux requêtes pourraient être imprécises (intervalle de valeurs), incertaines, voire les deux (par exemple, sous forme d'ensembles aléatoires).

Nous discutons ici du premier cas : les informations fournies sont des intervalles $R_q \ni \boldsymbol{X}_q$. De tels éléments de réponse sont entachés d'incertitude épistémique [8], l'expert ne donnant pas

de valeur précise pour $\boldsymbol{X}_q$. La cause de telles réponses peut être le processus d'élicitation lui-même, par exemple si l'on demande à l'expert de comparer la variable d'intérêt à un seuil.

5.2 Prise en compte de réponses imprécises

Stratégie naïve La stratégie naïve peut toujours être déployée, les distributions n'étant pas mises à jour en fonction des réponses.

Il reste possible de prendre une décision en calculant les probabilités a posteriori des classes, après avoir marginalisé les distributions conditionnelles préalablement *tronquées* par rapport aux intervalles R_q. Notons que dans ce cas, la stratégie de décision (1) pourrait être remplacée par un critère prudent. la réponse imprécise définit un ensemble crédal de probabilités a posteriori : pour toute classe $\omega_k \in \Omega$,

$$\mathcal{P}_k = \{\Pr(\omega_k|\boldsymbol{x}_q, \boldsymbol{x}_O) \text{ t.q. } \boldsymbol{x}_q \in R_q\}\,.$$

Lorsque cet ensemble crédal est spécifié, il est donc possible de calculer une entropie "pessimiste" pour faire de nouvelles requêtes (sur la variable $\boldsymbol{X}_q$ ou sur d'autres). Les critères d'entropie dans le cas de distributions mal connues [1, 2] peuvent être utilisés. Il est également possible de prendre une décision en utilisant des stratégies prudentes [16] comme la dominance par intervalles.

Stratégie séquentielle Dans ce cas, la distribution de la variable observée partiellement peut être tronquée et la variable peut être requêtée à nouveau, comme mentionné ci-dessus. Le problème se situe au niveau de la mise à jour des distributions des autres variables. en effet, pour chaque classe ω_k, on sait que $\boldsymbol{X}_M|R_q, \boldsymbol{x}_O$ appartient à l'ensemble

$$\mathcal{F}_{\boldsymbol{X}_M|\omega_k,R_q\boldsymbol{x}_O} = \left\{f_{\boldsymbol{X}_M|Z=\omega_k,\boldsymbol{x}_q,\boldsymbol{x}_O} \text{ t.q. } \boldsymbol{x}_q \in R_q\right\}.$$

Remarquons que pour tout $X_j \in \boldsymbol{X}_M$,

$$f_{X_j|\omega_k,\boldsymbol{x}_O,R_q}(x_j) = \frac{\Pr(R_q|\omega_k, \boldsymbol{x}_O, x_j) f_{X_j|\omega_k,\boldsymbol{x}_O}(x_j)}{\Pr(R_q|\omega_k, \boldsymbol{x}_O)}. \quad (6)$$

Dans cette expression, le dénominateur et le terme de droite du numérateur se calculent facilement, mais $\Pr(R_q|\omega_k, \boldsymbol{x}_O, x_j)$ dépend de x_j. La mise à jour des distributions conditionnelles en fonction des réponses aux requêtes, ou l'estimation de l'entropie, deviennent donc des problèmes difficiles, même dans le cas gaussien. Notons que des travaux récents concernant la propagation de l'incertitude via des copules [17] pourrait constituer un angle d'attaque.

6 Conclusion et perspectives

Résumé En classification supervisée, nous étudions le cas où les exemples de test à classer ne sont pas observés, ou seulement partiellement. Un oracle peut être interrogé à propos des valeurs manquantes dans le but de compléter les exemples avant de prendre une décision. Nous proposons d'exploiter les distributions des classes (estimées), et de choisir les variables avec le pouvoir discriminant le plus élevé au moyen d'un critère d'entropie. Une approche naïve consiste à calculer l'entropie pour chaque variable une fois pour toutes ; une autre, séquentielle, consiste à mettre à jour les distributions en fonction des réponses de l'oracle avant de recalculer l'entropie et choisir une nouvelle variable, l'enchaînement des questions dépendant alors de l'exemple à classer.

Les deux stratégies sont évaluées sur divers jeux de données avec un modèle d'analyse discriminante quadratique (hypothèse gaussienne dans chaque classe). Les expériences montrent qu'il peut être fructueux de choisir les variables à compléter. La procédure séquentielle semble pouvoir donner de meilleurs résultats que la stratégie naïve, mais se montre bien moins "distributionnellement robuste".

Nous discutons ensuite du cas où l'expert interrogé peut donner des réponses imprécises, sous la forme d'ensembles de valeurs possibles pour la variable requêtée. Bien que l'approche naïve puisse encore être utilisée dans ce cas, le problème de la mise à jour des distributions conditionnelles constitue un obstacle au

déploiement de l'approche séquentielle.

Perspectives Ce travail préliminaire ouvre de nombreuses perspectives de recherche. Une première direction est d'explorer plus avant le cas de réponses imprécises, en spécifiant par exemple des ensembles de distributions et en en déduisant des ensembles de probabilités a posteriori, à partir desquelles il serait possible d'utiliser une extension de l'entropie [1, 2] ou des stratégies de décision robustes [16, 6]. Il reste également à étudier le cas de réponses imprécises et incertaines.

La question de la robustesse à l'hypothèse de distribution faite se pose également. L'utilisation d'ensembles de distributions dans chaque classe pourrait apporter une réponse à ce problème, particulièrement critique dans le cas de l'approche séquentielle (les distributions conditionnelles classes pouvant s'éloigner fortement des vraies suite aux conditionnements successifs résultant des réponses de l'oracle).

La question du nombre de requêtes à formuler reste ouverte. Les expériences suggèrent qu'il est possible d'atteindre un très bon taux de performance à partir d'un nombre restreint de requêtes bien choisies. Il semble intéressant de quantifier la quantité (espérée) d'information restant à identifier. Cependant, bien qu'il existe une règle de chaînage pour l'entropie, cette propriété n'est plus valable lorsque les distributions des classes sont mises à jour par conditionnements successifs. Nous pourrions également étendre le problème du choix du nombre de requêtes lorsque ces dernières ont un coût (pouvant dépendre de l'exemple considéré), via un rapport coût-bénéfice.

D'autres pistes semblent intéressantes. La question de la complexité des deux stratégies (qui ne pouvait faire l'objet d'une étude détaillée dans cet article, faute de place) est également centrale dès lors qu'il s'agit de choisir entre les approches naïve et séquentielle. Par ailleurs, nous avons mentionné qu'il semble exister des séquences typiques des classes. Il pourrait être possible de pré-calculer de telles séquences, par exemple en utilisant un ensemble de validation, de manière à réduire le coût de l'approche séquentielle. Il pourrait également être intéressant d'explorer d'autres critères de choix des requêtes — ce problème restant fortement corrélé à celui de la robustesse de l'approche de requêtage.

Références

[1] J. Abellán and S. Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 11(5) :587–598, 2005.

[2] J. Abellán and S. Moral. Upper entropy of credal sets. applications to credal classification. *International Journal of Approximate Reasoning*, 39(2-3) :235–255, 2005.

[3] Alessandro Antonucci, Giorgio Corani, and Sandra Gabaglio. Active learning by the naive credal classifier. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, pages 3–10, 2012.

[4] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[5] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15 :201–221, 1994.

[6] Thierry Denœux. Decision-making with belief functions : a review. *International Journal of Approximate Reasoning*, 109 :87–110, 2019.

[7] D. Dua and C. Graff. UCI Machine Learning repository, 2019.

[8] S. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering and System Safety*, 54(2-3) :217–223, 1996.

[9] Pallika Kanani and Prem Melville. Prediction-time active feature-value

acquisition for customer targeting. In *Advances in Neural Information Processing Systems*, 2008.

[10] Edwin Lughofer. Hybrid active learning for reducing the annotation effort of operators in classification systems. *Pattern Recognition*, 45(2) :884–896, 2012.

[11] David MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[12] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers : a comparison of logistic regression and naive bayes. In *NIPS'01 : Proceedings of the 14th International Conference on Neural Information Processing Systems*, pages 841–848, 2001.

[13] Christian P. Robert. *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer, 2007.

[14] Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4) :664–684, 2009.

[15] Maytal Saar-Tsechansky and Foster Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8 :1625–1657, 2007.

[16] M.C.M. Troffaes. Decision making under uncertainty using imprecise probabilities. *Int. J. of Approximate Reasoning*, 45 :17–29, 2007.

[17] Jiaxin Zhang and Michael Shields. On the quantification and efficient propagation of imprecise probabilities with copula dependence. *International Journal of Approximate Reasoning*, 122 :24–46, 2020.