Classification supervisée pour de grands nombres de classes à prédire : une approche par co-partitionnement des variables explicatives et à expliquer

Marc Boullé *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Dans la phase de préparation des données du data mining, les méthodes de discrétisation et de groupement de valeurs supervisé possèdent de nombreuses applications : interprétation, estimation de densité conditionnelle, sélection de type filtre des variables, recodage des variables en amont des classifieurs. Ces méthodes supposent habituellement un faible nombre de valeur à expliquer (classes), typiquement moins d'une dizaine, et trouvent leur limite quand leur nombre augmente. Dans cet article, nous introduisons une extension des méthodes de discrétisation et groupement de valeurs, consistant à partitionner d'une part la variable explicative, d'autre part la variable à expliquer. Le meilleur co-partitionnement est recherché au moyen d'une approche Bayesienne de la sélection de modèle. Nous présentons ensuite comment utiliser cette méthode de prétraitement en préparation pour le classifieur Bayesien naïf. Des expérimentations intensives démontrent l'apport de la méthode dans le cas de centaines de classes.

1 Introduction

L'objectif de la classification supervisée est de prédire la valeur d'une variable catégorielle à expliquer connaissant l'ensemble des valeurs des variables explicatives, numériques ou catégorielles. La plupart des problèmes de classification considérés usuellement se limitent à la prédiction d'une valeur booléenne, ou d'une variable comportant un nombre très faible de valeurs, typiquement moins d'une dizaine. On rencontre néanmoins des problèmes où ce nombre de valeurs à expliquer est plus important, comme par exemple la reconnaissance de chiffres manuscrits, la reconnaissance de caractères ou la classification de textes. Les applications émergentes de ciblage publicitaire sur internet sont amenées à considérer le cas du choix d'un bandeau publicitaire parmi plusieurs centaines pour maximiser le taux de clic lors de la navigation des internautes. Les méthodes existantes supposent au moins implicitement un faible nombre de classes, et sont potentiellement moins performantes dans le cas de grands nombres de classes, avec peu d'individus par classe Il s'agit ici d'envisager le problème de classification dans son cadre le plus général sans faire l'hypothèse d'un nombre restreint de

classes. On s'intéresse ici à la préparation des données univariée consistant à discrétiser les variable numériques et grouper les valeurs des variables catégorielles Ces méthodes ont été largement traitées dans le bibliographie, en prétraitement pour les arbres de décision (Breiman et al., 1984; Quinlan, 1993; Zighed et Rakotomalala, 2000) ou pour le classifieur Bayesien naïf (Dougherty et al., 1995; Liu et al., 2002; Yang et Webb, 2002). L'objectif de cet article est d'étendre la préparation des données au cas de nombreuses classes.

Les méthodes de prétraitement par discrétisation consistent à partitionner la variable explicative, en intervalles dans le cas numériques et en groupes de valeurs dans le cas catégoriel, de façon à obtenir une estimation de la probabilité conditionnelle des classes à prédire. Quand le nombre de classes est faible, ces méthodes fournissent une estimation robuste, mais quand ce nombre augmente, ces méthodes sont soit sujettes au sur-apprentissage, soit contraintes de sous-partitionner la variable explicative. Une solution est alors de considérer le partitionnement joint de la variable explicative et de la variable à expliquer. La méthode heuristique présentée dans (Ritschard et al., 2001) s'intéresse au problème du groupement des lignes et colonnes d'un tableau de contingence et maximise un critère d'association, tel le coefficient V de Cramer, T de Tschuprow ou ϕ de Pearson. L'algorithme proposé est en O(N^5) où N est le nombre d'individus, ce qui le limite au cas de variables ayant de faibles nombres de valeurs. Dans (Nadif et Govaert, 2005), le problème est posé sous la forme d'un modèle de mélange par bloc, et optimisé au moyen de l'algorithme EM (expectation-maximisation). Cette approche est adaptée à l'analyse exploratoire dans le cadre d'un coclustering individus x variables, mais en raison de son temps de calcul, elle n'est pas adaptée à la préparation des données avec potentiellement de nombreuses variables explicatives à traiter. Parmi les méthodes apparentées, on peut également citer les approches de type ECOC (Error-correcting output codes) (Dietterich et Bakiri, 1995) qui permettent d'appliquer un classifieur binaire dans le cas multi-classes, en réduisant le problème multi-classes à une série de problèmes binaires, basées sur des bi-partitions des classes. L'objectif de notre approche est non pas de permettre l'utilisation de classifieurs binaires dans le cas multi-classes, mais d'améliorer la précision et la robustesse des estimateurs de densité conditionnelle univariés, en recherchant pour chaque variable explicative la partition des classes la plus adaptée, potentiellement différente par variable explicative.

Dans cet article, nous étendons l'approche MODL utilisée dans le cas de la discrétisation supervisée (Boullé, 2006) et du groupement de valeurs supervisé (Boullé, 2005). Cette approche pose le problème du prétraitement univarié comme un problème de sélection de modèles, un modèle étant défini par une partition des valeurs explicatives en intervalles ou groupes de valeurs et une distribution multinomiale des classes dans chaque partie. Le modèle de prétraitement est ici étendu en partitionnant conjointement les classes, et en se limitant à une distribution multinomiale des groupes de classes dans chaque partie explicative. Il s'agit alors de trouver un compromis entre les modèles fortement discriminants, basés sur des groupes de classes de faible cardinalité, et les modèles plus robustes mais moins discriminants, exploitant des groupes de classes de forte cardinalité. Ce compromis est trouvé en recherchant le meilleur modèle selon une approche Bayesienne.

L'article est organisé de la façon suivante. La partie 2 rappelle l'approche MODL utilisée pour les méthodes de prétraitement supervisé univarié. La partie 3 introduit l'extension de cette approche au cas des variables à expliquer comportant de grands nombres de classes. La partie 4 présente l'impact des ces prétraitements étendus pour le classifieur Bayesien naïf. La partie 5 évalue les performances de la méthode. Enfin, la partie 6 conclut cet article.

2 Prétraitements supervisés MODL

Cette section rappelle les principes de l'approche MODL[1] dans le cas de la discrétisation supervisée (Boullé, 2006) et du groupement de valeurs supervisé (Boullé, 2005).

2.1 Discrétisation supervisée

La discrétisation supervisée traite des variables explicatives numériques. Elle consiste à partitionner la variable explicative en intervalles, en conservant le maximum d'information relative aux classes. Un compromis doit être trouvé entre la finesse de l'information prédictive, qui permet une discrimination efficace des classes, et la fiabilité statistique, qui permet une généralisation du modèle de discrétisation. Dans l'approche MODL, la discrétisation supervisée est formulée en un problème de sélection de modèles. Une approche Bayésienne est appliquée pour choisir le meilleur modèle de discrétisation, qui est recherché en maximisant la probabilité $p(Model|Data)$ du modèle sachant les données. En utilisant la règle de Bayes, et puisque la quantité $p(Data)$ est constante pour un même jeu de données en apprentissage, il s'agit alors de maximiser $p(Model)p(Data|Model)$, c'est-à-dire un terme d'a priori sur les modèles et un terme de vraisemblance des données connaissant le modèle.

Dans un premier temps, une famille de modèles de discrétisation est explicitement définie. Les paramètres d'une discrétisation particulière sont le nombre d'intervalles, les bornes des intervalles et les effectifs des classes par intervalle. Dans un second temps, une distribution a priori est proposée pour cette famille de modèles. Cette distribution a priori exploite la hiérarchie des paramètres : le nombre d'intervalles est d'abord choisi, puis les bornes des intervalles et enfin les effectifs par classe. Le choix est uniforme à chaque étage de cette hiérarchie. De plus, les distributions des classes par intervalle sont supposées indépendantes entre elles.

Soient N le nombre d'individus, J le nombre de classes, I le nombre d'intervalles, $N_{i.}$ le nombre d'individus dans l'intervalle i et N_{ij} le nombre d'individus de la classe j dans l'intervalle i. Dans le contexte de la classification supervisée, les nombre d'individus N et de classes J sont supposés connus. Un modèle de discrétisation supervisée est entièrement caractérisé par les paramètres $\{I, \{N_{i.}\}_{1\leq i\leq I}, \{N_{ij}\}_{1\leq i\leq I, 1\leq i\leq J}\}$.

En utilisant la définition de la famille de modèles de discrétisation et de sa distribution a priori, la formule de Bayes permet de calculer explicitement les probabilités a posteriori des modèles connaissant les données. En prenant le log négatif de ces probabilités, cela conduit au critère d'évaluation fourni dans la formule (1).

$$\log N + \log \binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log \binom{N_{i.}+J+1}{J-1} + \sum_{i=1}^{I} \frac{N_{i.}!}{N_{i1}!N_{i2}!...N_{iJ}!} \qquad (1)$$

Les trois premiers termes représentent la probabilité a priori du modèle : choix du nombre d'intervalles, des bornes des intervalles, et de la distribution des classes dans chaque intervalle. Le dernier terme représente la vraisemblance, c'est-à-dire la probabilité d'observer les classes connaissant le modèle de discrétisation.

Une discrétisation quasi-optimale est recherchée en optimisant le critère d'évaluation, au moyen de l'heuristique gloutonne ascendante décrite dans (Boullé, 2006). A l'issue de cet

[1] Outil disponible en shareware sur http ://perso.rd.francetelecom.fr/boulle/

algorithme d'optimisation, des post-optimisations sont effectuées au voisinage de la meilleure solution, en évaluant des combinaisons de coupures et de fusions d'intervalles. L'algorithme exploite la décomposabilité du critère sur les intervalles pour permettre après optimisation de se ramener à une complexité algorithmique en O($JN \log N$).

2.2 Groupement de valeurs supervisé

Le cas des variables explicatives catégorielles est traité au moyen d'une approche similaire, en évaluant les modèles de groupement de valeurs. Dans le cas numérique, il s'agit de partitionner les valeurs explicatives, avec une contrainte d'adjacence entre valeurs (partitionnement en intervalles). Dans le cas catégoriel, il s'agit toujours de partitionner les valeurs explicatives, cette fois sans aucune contrainte (partitionnement en groupes de valeurs). Soient N le nombre d'individus, V le nombre de valeurs explicatives, J le nombre de classes, I le nombre de groupes de valeurs, $N_{i.}$ le nombre d'individus dans le groupe de valeur i et N_{ij} le nombre d'individus de la classe j dans le groupe i. L'application de l'approche Bayesienne de la sélection de modèle conduit ici à un critère d'évaluation d'un groupement de valeurs, fourni dans la formule (2). Cette formule possède une structure similaire à celle de la formule (1), en remplaçant dans les deux premiers termes la probabilité a priori d'une partition en intervalles par celle d'une partition en groupes de valeurs.

$$\log V + \log B\left(V, I\right) + \sum_{i=1}^{I} \log \binom{N_{i.} + J - 1}{J - 1} + \sum_{i=1}^{I} \log \frac{N_{i.}!}{N_{i1}! N_{i2}! \ldots N_{iJ}!}. \qquad (2)$$

$B(V, I)$ est le nombre de répartitions des V valeurs explicatives en I groupes (éventuellement vides). Pour $I = V$, $B(V, I)$ correspond au nombre de Bell. Dans le cas général, $B(V, I)$ peut s'écrire comme une somme de nombre de Stirling de deuxième espèce (nombre de partitions de V valeur en i groupes non vides) : $B(V, I) = \sum_{i=1}^{I} S(V, i)$. Le critère d'évaluation des groupements de valeurs est optimisé au moyen d'une heuristique gloutonne ascendante décrite dans (Boullé, 2005). Des étapes de pré-optimisation et post-optimisation sont utilisées, de façon à garantir une complexité algorithmique en O($JN \log N$) sans sacrifier aux performances de la méthode.

3 Généralisation aux grands nombres de classes à prédire

En présence d'un grand nombre de classes, il n'est pas raisonnable de modéliser directement la distribution de ces classes, en raison des effectifs par classe potentiellement très faibles. On propose de partitionner les classes en groupes de classes, ce qui permet de se ramener à un problème de classification supervisée standard portant sur un faible nombre de groupes de classes, puis de décrire la classe effective de chaque individu localement à son groupe de classes.

Soit Y une variable catégorielle à expliquer comportant W classes. Il s'agit d'étendre les modèles de prétraitement supervisé en incorporant un groupement des W classes en J groupes. Le cas standard peut être considéré comme un cas particulier, pour lequel $J = W$. Ici, W est supposé connu à l'avance alors que le nombre J de groupes est un paramètre à estimer.

Notations

– N : nombre d'individus de l'échantillon
– Y : variable catégorielle à expliquer
– W : nombre de classes de la variable à expliquer (connu)
– J : nombre de groupes de classes de la variable à expliquer (inconnu)
– $j(w)$: index du groupe auquel est rattaché la valeur à expliquer w
– $N_{.j}$: nombre d'individus du groupe à expliquer j
– m_j : nombre de classes du groupe j
– n_w : nombre d'individus pour la classe w

Pour un nombre de groupes J fixé, il s'agit de décrire une partition des W classes en J groupes, ce qui revient à spécifier $\{j(w)\}_{1 \leq w \leq W}$. De façon similaire au cas du groupement des valeurs d'une variable explicative décrit en section 2.2, la spécification du groupement des classes aboutit à l'ajout des nouveaux termes d'a priori suivants :

$$\log W + \log B(W, J). \tag{3}$$

Notons qu'une fois cette partition spécifiée, les nombres m_j de valeurs par groupe s'en déduisent et ne font donc pas partie du paramétrage de modélisation.

On se ramène ensuite au cas classique du prétraitement univarié supervisé présenté en section 2. Les modèles de partitionnement de la variable explicative sont exploités pour définir dans chaque partie explicative la distribution des individus sur les J groupes de classes. L'effectif par groupe $N_{.j}$ est déduit par sommation des effectifs par groupe de classes sur l'ensemble des I parties explicatives, selon $N_{.j} = \sum_{i=1}^{I} N_{ij}$.

Chaque individu étant associé à un groupe de classes, il s'agit désormais de préciser à quelle classe spécifique il est associé. Pour ce faire, on décrit localement à chaque groupe j la distribution des individus du groupe sur les classes du groupe, au moyen d'un modèle multinomial de distribution des N_j individus du groupe sur ses m_j classes. Comme précédemment, on utilise un a priori uniforme pour le paramétrage de ce modèle multinomial, ce qui conduit pour chaque groupe à l'ajout du nouveau terme d'a priori suivant :

$$\log \binom{N_{.j} + m_j - 1}{m_j - 1} \tag{4}$$

La vraisemblance de la distribution des individus sur les groupes est gérée par le modèle de prétraitement standard. Il faut ici ajouter un terme de vraisemblance localement à chaque groupe pour la distribution des individus du groupe sur les classes du groupe, au moyen d'un terme du multinôme :

$$\log N_{.j}! - \sum_{\{w; j(w)=j\}} \log n_w! \tag{5}$$

En sommant sur l'ensemble des groupes à expliquer, on obtient

$$\log W + \log B(W, J) + \sum_{j=1}^{J} \log \binom{N_{.j} + m_j - 1}{m_j - 1} \tag{6}$$

pour les termes d'a priori, et

$$\sum_{j=1}^{J} \log N_{.j}! - \sum_{v=1}^{V} \log n_v! \tag{7}$$

pour les termes de vraisemblance.

Il s'agit maintenant de rajouter ces nouveaux termes d'a priori et de vraisemblance aux critères de prétraitement présentés en section 2. Dans le cas de la discrétisation supervisée par exemple, la formule (1) est étendue en :

$$\begin{aligned} \log N + \log \binom{N+I-1}{I-1} + \quad & \sum_{i=1}^{I} \log \binom{N_{i.}+J+1}{J-1} \quad + \sum_{i=1}^{I} \frac{N_{i.}!}{N_{i12}!N_{i2}!...N_{iJ}!} \\ + \log W + \log B(W,J) + \quad & \sum_{j=1}^{J} \log \binom{N_{.j}+m_j-1}{m_j-1} \quad + \sum_{j=1}^{J} \log N_{.j}! - \sum_{v=1}^{V} \log n_v! \end{aligned} \tag{8}$$

Le problème du partitionnement joint des variables explicatives et à expliquer s'apparente à celui de la discrétisation supervisée de deux variables explicatives, ce qui permet de réutiliser les mêmes algorithmes (bien que les critères soient sensiblement différent, ils possèdent la même structure). Dans cet article, nous avons utilisé les heuristiques décrites dans (Boullé, 2009), qui présentent l'avantage de la tenue de charge avec une complexité algorithmique en $O(N\sqrt{N}\log N)$, indépendante du nombre de valeurs par variable.

4 Impact sur le classifieur Bayesien naïf

Cette partie rappelle les principes du classifieur Bayesien naïf, puis détaille la prise en compte des prétraitements introduits en section 3 pour le calcul des scores de prédiction.

4.1 Le classifieur Bayesien Naïf

Soient $X = (X_1, X_2, \ldots, X_K)$ un ensemble de K variables explicatives numériques ou catégorielles et Y une variable catégorielle à expliquer, comportant W classes $\lambda_1, \lambda_2, \ldots \lambda_W$. Soit $x = (x_1, x_2, \ldots, x_K)$ les valeurs explicatives d'un nouvel individu à classer.

Le classifieur Bayesien associe à chaque individu la classe maximisant la probabilité conditionnelle a posteriori

$$P(Y = \lambda_w | X = x) \quad = \quad \frac{P(Y = \lambda_w) P(X = x | Y = \lambda_w)}{P(X = x)}.$$

Le classifieur Bayesien est optimal, mais il n'est pas calculable en pratique, puisqu'il suppose que l'on connaisse parfaitement la distribution jointe de probabilité conditionnelle. Le modèle Bayesien naïf (Langley et al., 1992) relâche fortement la contrainte sur l'estimation multivariée de probabilité conditionnelle, en faisant l'hypothèse *naïve* d'indépendance des variables explicatives conditionnellement à la variable à expliquer. Parfois qualifié d'*idiot de Bayes* dans la littérature, le classifieur Bayesien naïf est souvent performant en pratique sur de nombreux jeux de données réels (Hand et Yu, 2001). Il est simple à mettre en œuvre, rapide à apprendre et à déployer, et peu sujet au sur-apprentissage, puisque l'espace des modèles est réduit à un singleton. En appliquant cette hypothèse naïve d'indépendance conditionelle des variables explicatives, on aboutit à :

$$P(Y = \lambda_w | X = x) \quad = \quad \frac{P(Y = \lambda_w) \prod_{k=1}^{K} P(X_k = x_k | Y = \lambda_w)}{P(X = x)}. \tag{9}$$

L'équation (9) est suffisante pour obtenir la classe la plus probable connaissant les variables descriptives. Dans les problèmes où un score de prédiction est nécessaire, la probabilité conditionnelle de la classe peut être obtenue par sommation au dénominateur sur les classes :

$$P(Y = \lambda_w | X = x) = \frac{P(Y = \lambda_w) \prod_{k=1}^{K} P(X_k = x_k | Y = \lambda_w)}{\sum_{v=1}^{W} P(Y = \lambda_v) \prod_{k=1}^{K} P(X_k = x_k | Y = \lambda_v)}. \quad (10)$$

4.2 Prise en compte des prétraitements univariés

A l'issue des prétraitements, chaque variable X_k est partitionnée en I_k parties explicatives (intervalles ou groupes de valeurs) pour l'estimation conditionnelle de Y, elle même partitionnées en J_k groupes de classes. Soient $N^k_{i_k.}$ l'effectif en apprentissage de la partie i_k de X_k, $N^k_{.j_k}$ celui de la partie j_k de Y et $N^k_{i_k j_k}$ celui de la cellule (i_k, j_k).

En exploitant le prétraitement par partitionnement joint de X_k et Y, soient $P^k_{i_k}(x_k)$ la partie à laquelle la valeur explicative x_k appartient et $G^k_{j_k}(\lambda_w)$ le groupe associé à la classe λ_w. Le modèle de prétraitement permet d'estimer les probabilités conditionnelles de façon constante par morceau, ce qui conduit à :

$$\begin{aligned} P(X_k = x_k | Y = \lambda_w) &= P\left(x_k \in P^k_{i_k}(x_k) | \lambda_w \in G^k_{j_k}(\lambda_w)\right), \\ P(X_k = x_k | Y = \lambda_w) &= \frac{N^k_{i_k j_k}}{N^k_{.j_k}}. \end{aligned} \quad (11)$$

Les probabilités a priori des classes peuvent elles être estimées par leur probabilité empirique $P(Y = \lambda_w) = n_w/N$ sur la base de l'effectif n_w de la classe λ_w et de la taille N de l'échantillon d'apprentissage. En exploitant ces estimations empiriques de probabilité, la formule (10) se calcule de la façon suivante :

$$P(Y = \lambda_w | X = x) = \frac{n_w \prod_{k=1}^{K} \frac{N^k_{i_k j_k}}{N^k_{.j_k}}}{\sum_{v=1}^{W} n_v \prod_{k=1}^{K} \frac{N^k_{i_k j_k}}{N^k_{.j_k}}}. \quad (12)$$

Afin d'éviter les probabilités nulles, les probabilités conditionnelles sont estimées en utilisant un m-estimate $(support + mp) = (coverage + m)$ avec $m = W/N$ et $p = 1/W$.

5 Expérimentation

Dans cette section, nous évaluons l'impact de notre méthode de prétraitement sur la préparation des données et la modélisation par le classifieur Bayesien naïf.

5.1 Exemple illustratif

A titre illustratif, nous utilisons le jeu de données Letter de l'UCI (Asuncion et Newman, 2007), qui consiste a prédire une lettre capitale (parmi 26) à partir d'une matrice de pixels en noir et blanc. Les 16 variables numériques descriptives sont des mesures concernant la taille

de la boîte contenant la lettre et des moments statistiques résumant la position des pixels noirs dans la boîte. A titre d'exemple, la largeur de la boîte (with) est une de ces mesures. La figure 1 présente sous forme d'histogramme bivarié les résultats de prétraitement de partitionnement joint de la variable explicative width, en 10 intervalles de largeur, et de la variable à expliquer, en 8 groupes de lettres. La hauteur des histogrammes représente la probabilité conditionnelle d'observer une lettre appartenant un groupe de lettres, sachant que sa largeur est dans un intervalle de valeurs. Par exemple, pour les très petites largeurs ($width <= 0.5$), on a une probabilité conditionnelle de 100% d'observer la lettre I. A l'inverse, pour les très grandes largeurs de lettre ($width > 10.5$), on peut observer dans 60% des cas soit M soit W, et dans 40% des cas une lettre parmi X, N, K ou H. Les autres situations correspondent à des cas intermédiaires. Globalement, le prétraitement permet d'obtenir une estimation constante par morceau des probabilités conditionnelles, et cette estimation est la meilleure connaissant les données selon l'approche Bayesienne de sélection de modèles utilisée.

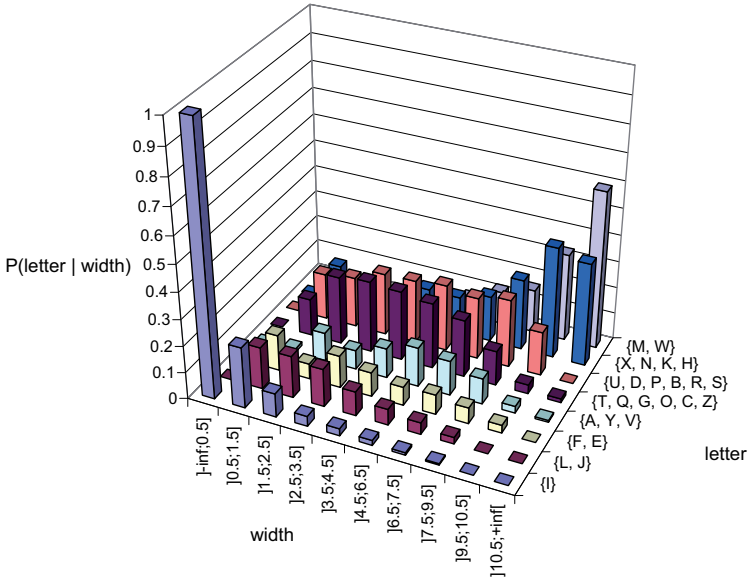

FIG. 1 – *Estimation des probabilités conditionnelles de la lettre à expliquer (letter) connaissant sa largeur (width) pour la base Letter.*

5.2 Expérimentation sur les bases de l'UCI

Afin d'étudier l'impact de la méthode de prétraitement avec groupement de classes sur le taux de bonne prédiction en test du classifieur Bayesien naïf (NB), nous avons étudié trois types de prétraitement :

– NB(G) : partitionnement supervisé des variables explicatives et à expliquer (section 3),
– NB : partitionnement supervisé des variables explicatives uniquement (section 2),
– nb : version standard, avec discrétisation non supervisée en 10 intervalles d'effectif égal dans le cas numérique et sans regroupement de valeurs dans le cas catégoriel.

Il est à noter que alors que chaque prétraitement univarié permet d'obtenir des estimations de probabilité conditionnelles par groupe de classes (cf. section 5.1), le prédicteur Bayesien naïf combine ces estimations correspondant à des partitions de classes potentiellement différentes pour obtenir des estimations de probabilité conditionnelle par classe (cf. formule 12).

Les expérimentations sont menées en utilisant 18 jeux de données de l'UCI (Asuncion et Newman, 2007) décrits en table 1, représentant une grande diversité de domaines, de nombres d'individus (N), de variables explicatives numériques ou catégorielles (K) et comportant au moins trois classes (W) avec des distributions de classes parfois déséquilibrées (Maj. rappelle la fréquence de la classe majoritaire). Le taux de bonne prédiction en test est évalué au moyen d'une validation croisée stratifiée à 10 niveaux.

Dataset	N	K	W	Maj.	NB(G)	NB	nb
Abalone	4177	8	28	0.165	**0.262** $\pm$ 0.022	0.243 $\pm$ 0.028	0.225 $\pm$ 0.020
Flag	194	29	8	0.309	**0.646** $\pm$ 0.083	0.636 $\pm$ 0.070	0.640 $\pm$ 0.088
Glass	214	10	6	0.355	**0.953** $\pm$ 0.046	0.949 $\pm$ 0.048	0.921 $\pm$ 0.042
Iris	150	4	3	0.333	0.913 $\pm$ 0.085	0.920 $\pm$ 0.088	**0.947** $\pm$ 0.040
Led	1000	7	10	0.114	**0.747** $\pm$ 0.038	0.743 $\pm$ 0.032	0.743 $\pm$ 0.032
Led17	10000	24	10	0.107	**0.738** $\pm$ 0.011	0.732 $\pm$ 0.010	0.736 $\pm$ 0.013
Letter	20000	16	26	0.041	**0.747** $\pm$ 0.013	**0.747** $\pm$ 0.013	0.712 $\pm$ 0.012
PenDigits	10992	16	10	0.104	**0.885** $\pm$ 0.012	0.884 $\pm$ 0.010	0.871 $\pm$ 0.010
Phoneme	2254	256	5	0.260	**0.876** $\pm$ 0.023	0.872 $\pm$ 0.025	0.875 $\pm$ 0.023
Satimage	6435	36	6	0.238	0.822 $\pm$ 0.009	**0.823** $\pm$ 0.010	0.812 $\pm$ 0.012
Segmentation	2310	19	7	0.143	0.921 $\pm$ 0.013	**0.923** $\pm$ 0.012	0.899 $\pm$ 0.011
Shuttle	58000	9	7	0.786	0.998 $\pm$ 0.000	**0.999** $\pm$ 0.000	0.992 $\pm$ 0.001
Soybean	376	35	19	0.138	0.918 $\pm$ 0.056	0.926 $\pm$ 0.068	**0.928** $\pm$ 0.060
Thyroid	7200	21	3	0.926	**0.994** $\pm$ 0.002	**0.994** $\pm$ 0.001	0.956 $\pm$ 0.007
Vehicle	846	18	4	0.258	0.595 $\pm$ 0.036	**0.618** $\pm$ 0.031	0.611 $\pm$ 0.035
Waveform	5000	21	3	0.339	**0.811** $\pm$ 0.022	0.810 $\pm$ 0.019	0.808 $\pm$ 0.024
Wine	178	13	3	0.399	**0.983** $\pm$ 0.026	**0.983** $\pm$ 0.026	0.977 $\pm$ 0.028
Yeast	1484	9	10	0.312	**0.575** $\pm$ 0.050	**0.575** $\pm$ 0.046	0.344 $\pm$ 0.032
Moyenne					0.799	0.799	0.778
V/=/D						2/14/2	8/10/0

TAB. 1 – *Taux de bonne prédiction en test sur les bases de l'UCI*

Les moyennes et écart types des résultats par jeu de données sont présentés en table 1, ainsi que la moyenne sur l'ensemble des jeu de données (Moyenne) et le nombre de victoires et défaites significatives (V/=/D) évaluées au seuil de 95% au moyen d'un test de Student. Les résultats confirment l'apport important des méthodes de prétraitement supervisé, avec 8 victoires significative de la méthode NB(G) par rapport à la méthode standard nb. En revanche, les résultats des méthodes NB(G) et NB sont équivalents. Pour les bases de l'UCI comportant de l'ordre d'une dizaine de classes, la méthode de prétraitement avec groupement des classes est donc intéressante pour l'interprétation (cf. section 5.1), mais son impact en taux de bonne classification est négligeable.

5.3 Expérimentation avec de nombreuses classes

Afin d'étudier l'impact de notre méthode dans le cas de grands nombres de classes, nous avons utilisé la base Mushroom de l'UCI pour construire une série de 10 bases artificielles. La base Mushroom comporte 8416 individus et 23 variables catégorielles, dont les nombres de valeurs sont rappelés sur la figure 2.

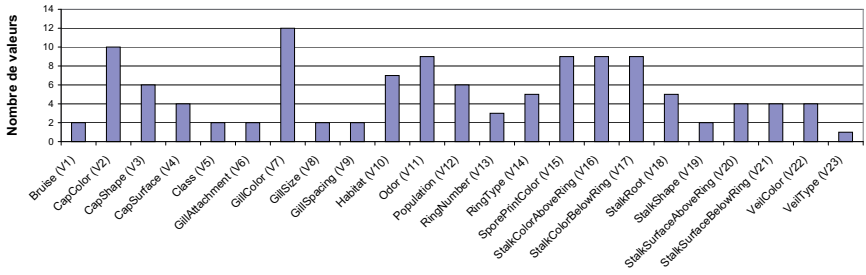

FIG. 2 – *Nombre de valeurs par variable de la base mushroom.*

Pour la première base, la variable de classe est la première variable (V1=Bruise), les 22 autres variables étant explicatives. Pour la seconde base, la variable de classe est le produit cartésien des deux premières variables (V1 x V2 = Bruise x CapColor), les 21 autres variables étant explicatives. Les bases suivantes sont construites de façon similaire, jusqu'à la dixième ayant pour variable de classe le produit cartésien des dix première variables, les 13 dernières étant explicatives. La table 2 présente les caractéristiques de ces dix base artificielles, avec des nombres de classes allant de 2 à 782 pour des classes majoritaires de 59.0% à 0.6%.

Target variable	Clas.	Maj.	NB(M)	NB
V1	2	0.599	**0.972** $\pm$ 0.004	**0.972** $\pm$ 0.004
V1xV2	17	0.174	**0.420** $\pm$ 0.006	0.418 $\pm$ 0.007
V1vV2xV3	56	0.084	**0.182** $\pm$ 0.003	0.146 $\pm$ 0.004
V1xV2...xV4	125	0.043	**0.105** $\pm$ 0.002	0.060 $\pm$ 0.008
V1xV2...xV5	159	0.024	**0.104** $\pm$ 0.002	0.049 $\pm$ 0.005
V1xV2...xV6	172	0.024	**0.104** $\pm$ 0.001	0.045 $\pm$ 0.004
V1xV2...xV7	495	0.017	**0.035** $\pm$ 0.001	0.008 $\pm$ 0.002
V1xV2...xV8	526	0.017	**0.034** $\pm$ 0.000	0.004 $\pm$ 0.002
V1xV2...xV9	554	0.017	**0.035** $\pm$ 0.000	0.005 $\pm$ 0.002
V1xV2...xV10	782	0.006	**0.015** $\pm$ 0.001	0.000 $\pm$ 0.001

TAB. 2 – *Taux de bonne prédiction en test sur 10 bases articicielles.*

On compare les taux de prédiction en test de NB(G) et NB comme en section 5.2 au moyen d'une validation croisée stratifiée à 10 niveaux. Cette fois les différences de performance sont importantes et significatives (8 victoire significatives sur 10), d'autant plus que le nombre de classes augmente. La figure 3, présente ces différences, en apprentissage et en test. Les deux méthodes obtiennent des résultats similaires en apprentissage, mais ces performances se

dégradent significativement pour NB dès que l'on atteint une cinquantaine de classes, alors que le sur-apprentissage reste faible pour NB(G) même avec plus de 500 classes.

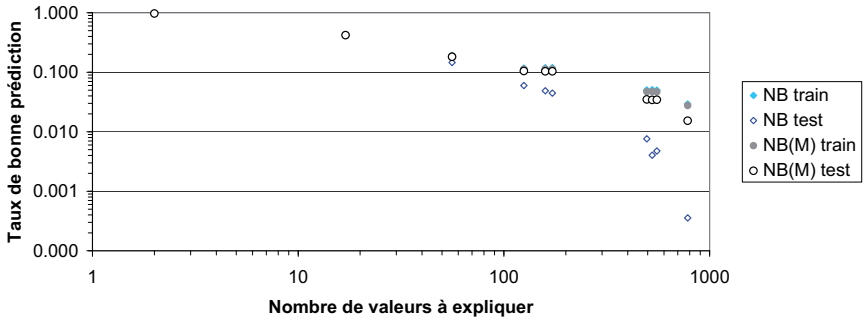

FIG. 3 – *Taux de bonne prédiction en apprentissage et test des méthodes NB(G) et NB, en fonction du nombre de valeurs à prédire.*

6 Conclusion

La méthode de prétraitement univarié supervisé présentée dans cet article se base sur un modèle de partitionnement joint, de la variable explicative en intervalles ou groupes de valeurs et de la variable à expliquer en groupe de classes. Ce partitionnement joint permet d'estimer de façon robuste la distribution de probabilité conditionnelle, sans être limité aux faibles nombres de classes. Le meilleur modèle de partitionnement joint est recherché au moyen d'une approche Bayesienne de la sélection de modèles. Des évaluations intensives sur les bases de l'UCI comportant de l'ordre d'une dizaine de classes, portant sur l'apport de la méthode en prétraitement du classifieur Bayesien naïf, montrent que la méthode obtient des performances équivalentes, mais non supérieures à celles de l'état de l'art. En revanche, quand le nombre de classes devient important, typiquement plusieurs centaines, les expérimentation démontrent un apport significatif de la méthode, avec une meilleure robustesse et des taux de bonne prédiction accrus. Ces résultats permettent d'étendre le champ d'application des méthodes de classification aux problèmes comportant de nombreuses classes. Des travaux futurs sont envisagés pour appliquer ce type d'approche au cas du ciblage publicitaire sur internet, pour lequel la classe à prédire, le bandeau publicitaire le plus susceptible d'entraîner un clic de l'internaute, peut comporter plusieurs centaines de valeurs.

Références

Asuncion, A. et D. Newman (2007). UCI machine learning repository.

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. (2009). Optimum simultaneous discretization with data grid models in supervised classification : a bayesian model selection approach. *Advances in Data Analysis and Classification 3*(1), 39–61.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *Classification and Regression Trees*. California : Wadsworth International.

Dietterich, T. et G. Bakiri (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research 2*, 263–286.

Dougherty, J., R. Kohavi, et M. Sahami (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 194–202. Morgan Kaufmann, San Francisco, CA.

Hand, D. et K. Yu (2001). Idiot bayes ? not so stupid after all ? *International Statistical Review 69*(3), 385–399.

Langley, P., W. Iba, et K. Thompson (1992). An analysis of Bayesian classifiers. In *10th National Conference on Artificial Intelligence*, pp. 223–228. AAAI Press.

Liu, H., F. Hussain, C. Tan, et M. Dash (2002). Discretization : An enabling technique. *Data Mining and Knowledge Discovery 4*(6), 393–423.

Nadif, M. et G. Govaert (2005). Block clustering of contingency table and mixture model. In *Advances in Intelligent Data Analysis VI*, Volume 3646 of *LNCS*, pp. 249–259. Springer.

Quinlan, J. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

Ritschard, G., D. A. Zighed, et N. Nicoloyannis (2001). Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. *Mathématiques et Sciences Humaines 154-155*, 81–98.

Yang, Y. et G. Webb (2002). A comparative study of discretization methods for naive-Bayes classifiers. In *Proceedings of the Pacific Rim Knowledge Acquisition Workshop*, pp. 159–173.

Zighed, D. et R. Rakotomalala (2000). *Graphes d'induction*. France : Hermes.

Summary

In the data preparation phase of data mining, supervised discretization and value grouping methods have numerous applications: interpretation, conditional density estimation, filter selection of input variables, variable recoding for classification methods. These methods usually assume a small number of output values, typically less than ten, and reach their limit when their number increases. In this paper, we extend discretization and value grouping methods, based on the partitioning of both the input and output variables. The best joint partitioning is searched by maximizing a Bayesian model selection criterion. We show how to exploit this preprocessing method as a preparation for the naïve Bayes classifier. Extensive experiments demonstrate the benefits of the approach in the case of hundred of output values.