Une méthode optimale d'évaluation bivariée pour la classification supervisée

Marc Boullé

France Télécom R&D, 2, avenue Pierre Marzin, 23300 Lannion
marc.boulle@orange-ft.com

Résumé. En préparation des données pour la classification supervisée, les méthodes filtres usuellement utilisées pour la sélection de variables sont efficaces en temps de calcul. Néanmoins, leur nature univariée ne permet pas de détecter les redondances ou les interactions constructives entre variables. Cet article présente une nouvelle méthode permettant d'évaluer l'importance prédictive jointe d'une paire de variables de façon automatique, rapide et fiable. Elle est basée sur un partitionnement de chaque variable exogène, en intervalles dans le cas numérique et groupes de valeurs dans le cas catégoriel. La grille de données exogène résultante permet alors d'évaluer la corrélation entre la paire de variables exogènes et la variable endogène. Le meilleur partitionnement bivarié est recherché au moyen d'une approche Bayesienne de la sélection de modèle. Les expérimentations démontrent les apports de la méthode, notamment une amélioration significative des performances en classification.

1 Introduction

Dans un projet de fouille de données, la phase de préparation des données vise à extraire une table de données pour la phase de modélisation (Pyle, 1999; Chapman et al, 2000). La préparation des données est non seulement coûteuse en temps d'étude, mais également critique pour la qualité des résultats escomptés. La préparation repose essentiellement sur la recherche d'une représentation pertinente pour le problème à modéliser, recherche qui se base sur une sélection de variables. L'objectif de la sélection de variable est triple: améliorer la performance prédictive des classifieurs, le temps d'apprentissage et de déploiement des modèles, et leur interprétabilité (Guyon et Elisseeff, 2003). Deux approches principales, filtre et enveloppe (Kohavi et John, 1997), ont été proposées dans la littérature. Les méthodes filtres évaluent la corrélation entre les variables exogènes et la variable endogène, indépendamment de la méthode de classification utilisée. Les méthodes enveloppes recherchent pour un modèle donné le meilleur sous-ensemble de variables. Les méthodes enveloppes, très coûteuses en temps de calcul, sont plutôt adaptées à la phase de modélisation. Parmi les méthodes filtres, les méthodes procédant par analyse univariée permettent d'ordonner les variables exogènes par importance prédictive décroissante. Elles sont classiquement utilisées en phase de préparation des données pour rapidement extraire un sous-ensemble de variables pertinent pour la modélisation à partir d'un ensemble de variables candidates potentiellement de grande taille. Dans cet article, nous nous focalisons sur l'approche filtre.

L'approche filtre la plus fréquemment utilisée repose sur la mise en œuvre de tests statistiques (Saporta, 1990), comme par exemple le test du Khi2 pour les variables exogènes catégorielles, ou les tests de Student ou de Fisher-Snedecor pour les variables exogènes numériques. Ces tests d'indépendance sont simples à mettre en œuvre, mais présentent de nombreux inconvénients. Ils se limitent à une discrimination entre variables dépendantes et indépendantes, sans permettre un ordonnancement précis des variables exogènes, et sont contraints par des hypothèses d'applicabilité fortes (effectifs minimaux, hypothèse de distribution gaussienne dans le cas numérique…). De nombreux autres critères d'évaluation de la dépendance entre deux variables ont été étudiés dans le contexte des arbres de décision (Zighed et Rakotomalala, 2000). Ces critères sont basés sur une partition de la variable exogène, en intervalles dans le cas numérique et en groupe de valeurs dans le cas catégoriel. En recherchant de façon non paramétrique un modèle de dépendance entre variables exogènes et endogène, ils permettent une évaluation fine de l'importance prédictive des variables exogènes. Dans le cas où tous les modèles de partitionnement de la variable exogène sont envisagés, un compromis doit être trouvé entre finesse de la partition et fiabilité statistique. Ce compromis est réalisé dans l'approche MODL (Boullé, 2005, 2006a) en formulant le problème comme un problème de sélection de modèle et en adoptant une approche Bayesienne.

Les méthodes filtres univariées restent néanmoins limitées, en étant aveugles aux interactions entre variables exogènes. Ainsi, les variables redondantes, apportant la même information, ne peuvent être détectées. De même, les variables exogènes qui seules sont non informatives et simultanément le sont (cas du XOR par exemple) ne sont pas détectables par les méthodes filtres. L'évaluation supervisée de l'importance d'une paire de variables exogènes, qui fait donc intervenir trois variables, a été peu étudiée dans la littérature. Les diagrammes de dispersion catégorisés par valeur endogène permettent une visualisation des paires de variables exogènes numériques, mais cela ne permet pas de quantifier l'information prédictive. Le regroupement simultané des lignes et des colonnes d'une table de contingence a été étudié dans un cadre général (Govaert et Nadif, 2006), ou dans le cadre des arbres de décision (Zighed et al, 2005) pour le partitionnement joint d'une variable exogène et de la variable endogène. Dans le cas de variables exogènes numériques, Muhlenbach et Rakotomalala (2002) utilisent un graphe de voisinage pour identifier des groupes d'instances ayant même valeur endogène, et projettent ces groupes sur les variables exogènes pour obtenir les bornes d'une discrétisation multivariée. Dans le cas non supervisé des règles d'association, Bay (2001) décrit une méthode de discrétisation multivariée permettant de mettre évidence les interactions entre variables numériques. Webb et al (2005) évaluent l'amélioration apportée au classifieur Bayesien naïf, en prenant en compte les paires de variables catégorielles, les probabilités conditionnelles étant estimées par comptage.

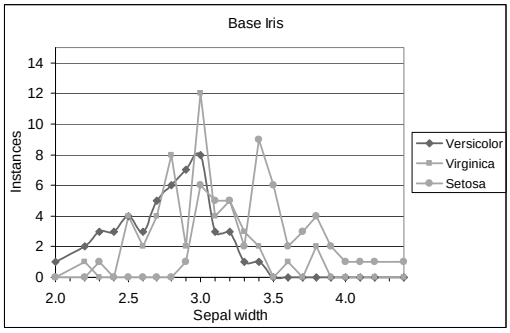
Nous proposons dans cet article une extension des méthodes de partitionnement univariées MODL au cas de l'analyse bivariée, pour tout type de paires de variables, numériques, catégorielles ou mixtes. Chaque variable est partitionnée, en intervalles ou groupe de valeurs selon son type, ce qui permet de distribuer les individus sur les cellules d'une grille bidimensionnelle. La corrélation entre la grille et la variable endogène est alors évaluée pour mesurer l'importance prédictive jointe de la paire de variables. Le compromis entre information et fiabilité, lié à la finesse de la grille, est établi au moyen d'une approche Bayesienne de la sélection de modèle, qui aboutit à un critère d'évaluation des partitionnements joints de variables exogènes. Ce critère d'évaluation est optimisé au moyen d'une heuristique réutilisant les techniques d'optimisation univariée MODL, en optimisant alternativement le partitionnement de chaque variable de la paire, le partitionnement de l'autre variable étant fixé.

L'article est organisé de la façon suivante. La section 2 rappelle l'approche MODL dans le cas univarié. La section 3 décrit l'extension de cette approche à l'analyse bivariée. La section 4 présente l'évaluation de la méthode. Enfin, la section 5 conclut cet article.

2 Analyse univariée supervisée

Cette section résume l'approche MODL de la discrétisation supervisée (Boullé, 2006a) et du groupement de valeurs supervisé (Boullé, 2005).

2.1 Discrétisation MODL



Versicolor	34	15	1
Viginica	21	24	5
Setosa	2	18	30
SepalWidth	]-inf;2.95[	[2.95;3.35[	[3.35;+inf[

FIG. 1 – *Discrétisation MODL de la variable largeur de sépale pour la classification de la base Iris en trois classes.*

La discrétisation supervisée traite des variables exogènes numériques. Elle consiste à partitionner la variable exogène en intervalles, en conservant le maximum d'information relative à la variable endogène. A titre illustratif, la figure 1 présente le nombre d'individus par classe du jeu de données Iris (Blake et Merz, 1996), pour chaque valeur de la variable largeur de sépale. Un compromis doit être trouvé entre la finesse de l'information prédictive, qui permet une discrimination efficace des valeurs endogènes, 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 modèle. Une approche Bayesienne 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 $p(Data)$ ne dépend pas du jeu de données, 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 endogènes 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 endogène. Le choix est uniforme à chaque

étage de cette hiérarchie. De plus, les distributions des valeurs endogènes par intervalle sont supposées indépendantes entre elles.

Soient N le nombre d'individus, J le nombre de classes endogènes, 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 supervisé est entièrement caractérisé par les paramètres $\{ I, \{N_{i.}\}_{1\le i\le I}, \{N_{ij}\}_{1\le i\le I, 1\le j\le J} \}$.

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).

$$\log(N) + \log\left(C_{N+I-1}^{I-1}\right) + \sum_{i=1}^{I} \log\left(C_{N_{i.}+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log\left(N_{i.}!/N_{i1}!N_{i2}!...N_{iJ}!\right). \quad (1)$$

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 valeurs endogènes dans chaque intervalle. Le dernier terme représente la vraisemblance, c'est à dire la probabilité d'observer les valeurs de la variable endogène connaissant le modèle de discrétisation.

La discrétisation optimale est recherchée en optimisant le critère d'évaluation, au moyen de l'heuristique gloutonne ascendante décrite dans (Boullé, 2006a). A l'issue de cet 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 optimisations de se ramener à une complexité algorithmique en O ($JN \log(N)$).

2.2 Groupement de valeurs MODL

Valeur	EDIBLE	POISONOUS	Effectif
BROWN	55.2%	44.8%	1610
GRAY	61.2%	38.8%	1458
RED	40.2%	59.8%	1066
YELLOW	38.4%	61.6%	743
WHITE	69.9%	30.1%	711
BUFF	30.3%	69.7%	122
PINK	39.6%	60.4%	101
CINNAMON	71.0%	29.0%	31
GREEN	100.0%	0.0%	13
PURPLE	100.0%	0.0%	10

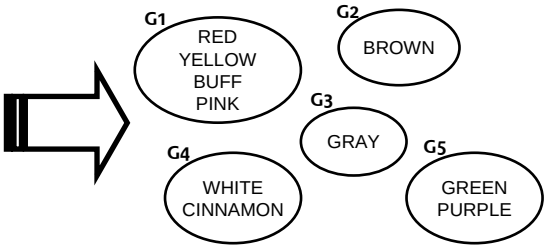


FIG. 2 – *Groupement de valeurs MODL de la variable couleur de chapeau pour la classification de la base Mushroom en deux classes.*

Le cas des variables exogènes 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 exogènes, avec une contrainte d'adjacence entre valeurs (partitionnement en intervalles). Dans le cas catégoriel, il s'agit toujours de partitionner les valeurs exogènes, cette fois sans aucune contrainte (partitionnement en groupes de valeurs). La figure 2 illustre le groupement des valeurs de la variable couleur de chapeau pour la classification de la base Mushroom (Blake et Merz, 1996).

Soient N le nombre d'individus, V le nombre de valeurs exogènes; J le nombre de classes endogènes, 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.

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

$B(V,I)$ est le nombre de répartitions des V valeurs exogènes en I groupes (éventuellement vides). Pour $I=V$, $B(V,I)$ correspond au nombre de Bell. Pour $I>V$, $B(V,I)$ s'écrit comme une somme de nombres de Stirling de deuxième espèce.

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.

Les méthodes de discrétisation et groupement de valeurs MODL produisent le partitionnement des valeurs exogènes le plus probable connaissant les données. Des évaluations comparatives intensives (Boullé, 2005, 2006a) ont mis en évidence les apports de la méthode, tant d'un point de vue informatif que prédictif.

3 Extension à l'analyse bivariée supervisée

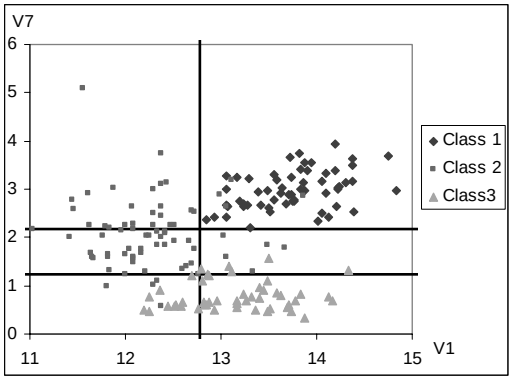
Nous présentons dans cette section la nouvelle méthode d'analyse bivariée, qui étend l'approche MODL à l'analyse supervisée des paires de variables exogènes. Après avoir introduit le problème au moyen d'un exemple illustratif, nous présentons un critère d'évaluation et un algorithme d'optimisation de ce critère.

3.1 Intérêt du partitionnement joint de deux variables

La figure 3 présente le diagramme de dispersion des variables V1 et V7 du jeu de données Wine (Blake et Merz, 1996), catégorisé par valeur endogène. Chaque variable isolément est faiblement discriminante. La variable V1 ne peut séparer les classes 1 et 3 au delà de la valeur 13. De même, la variable V7 confond les classes 1 et 2 au-delà de la valeur 2. Les deux variables conjointement autorisent une meilleure discrimination des classes.

L'approche utilisée pour qualifier l'information prédictive contenue dans la paire de variables repose sur un partitionnement des individus en une grille de données. Chaque variable exogène est partitionnée en intervalles (ou groupes de valeurs selon son type). Le produit cartésien des deux partitions univariées répartit les individus sur une grille de données, dont les cellules sont définies par des paires d'intervalles. Le lien avec la variable endogène se fait au moyen de la distribution des valeurs endogènes dans chaque cellule. Par exemple dans la figure 3, la variable V1 est discrétisée en 2 intervalles (borne 12.78) et la variable V7 en 3 intervalles (bornes 1.235 et 2.18). Les individus se répartissent dans les 6 cellules de la grille bidimensionnelle ainsi définie. Dans chaque cellule, nous obtenons une distribution des valeurs endogènes. Par exemple, le tableau de la figure 3 montre que 63 individus ont abouti

dans la cellule définie par les intervalles]12.78, +inf[sur V1 et]2.18, +inf[sur V7. Ces 63 individus sont distribués en 59 individus sur la classe 1 et 4 individus sur la classe 2.



V7xV1	]-inf;12.78]	]12.78;+inf[
]2.18;+inf[	(0, 23, 0)	(59, 0, 4)
]1.235;2.18]	(0, 35, 0)	(0, 5, 6)
]-inf;1.235]	(0, 4, 11)	(0, 0, 31)

FIG. 3 – *Diagramme de dispersion des variables exogènes V1 et V7 catégorisé par classe pour la base de données Wine. La discrétisation 2D optimale MODL est représentée sur le diagramme de dispersion et le triplet d'effectif par classe endogène est résumé par cellule de la grille de discrétisation 2D dans le tableau de droite.*

Les grilles de données sont d'autant plus fiables qu'elles contiennent plus d'individus par cellule, et d'autant plus informatives que les cellules permettent de bien discriminer les valeurs endogènes.

3.2 Critère d'évaluation

Nous utilisons ici la même approche que dans le cas univarié pour rechercher le meilleur compromis entre information et fiabilité, en introduisant une famille de modèles de partitionnement bivariés, puis en choisissant le meilleur modèle au moyen d'une approche Bayesienne. Nous nous intéressons d'abord au cas des variables exogènes numériques, avant de généraliser à tout type de paires de variables, numériques, catégorielles ou mixtes.

Définition 1. Un modèle de partitionnement bivarié supervisé est défini par une partition en intervalles pour chaque variable exogène, et par la distribution des valeurs endogènes pour chaque cellule de la grille de données déduite du croisement des deux partitions univariées.

Notations.

- X_1, X_2: variables exogènes,
- N: nombre d'individus,
- J: nombre de valeurs endogènes,
- I_1, I_2: nombre d'intervalles pour chaque variable exogène,
- $N_{i_1.}, N_{i_2.}$: nombres d'individus de l'intervalle i_1 (ou i_2) de la variable X_1 (ou X_2),
- $N_{i_1 i_2.}$: nombre d'individus de la cellule exogène (i_1, i_2) des variables (X_1, X_2),
- $N_{i_1 i_2 j}$: nombre d'individus de la cellule (i_1, i_2) pour la valeur endogène j.

Un modèle de partitionnement bivarié supervisé décrit la distribution des valeurs endogènes, connaissant les valeurs exogènes. Il est entièrement défini par les nombres d'intervalles (I_1, I_2), les bornes des intervalles ($N_{i_1.}, N_{i_2.}$) et la distribution des valeurs endogènes par cellule de la grille de données ($N_{i_1 i_2 j}$). Il est à noter que les nombres d'individus $N_{i_1 i_2.}$ par cellule de la grille ne font pas partie des paramètres du modèle: ils sont déduits du jeu de données à partir des partitionnements de chaque variable exogène.

Nous introduisons dans la définition 2 un a priori sur la distribution des paramètres des modèles de partitionnement bivarié supervisé, exploitant la hiérarchie des paramètres.

Définition 2. L'a priori hiérarchique sur l'espace des modèles de partitionnement bivarié supervisé est défini de la façon suivante:

- les nombres d'intervalles sont indépendants entre eux, et compris entre 1 et *N* de façon équiprobable,
- pour chaque variable exogène, pour un nombre d'intervalle donné, toutes les partitions en intervalles sur les rangs de la variable exogène sont équiprobables,
- pour chaque cellule de la grille de données, toutes les distributions des valeurs de la variable endogène sont équiprobables,
- les distributions des valeurs de la variable endogène sur chaque cellule sont indépendantes entre elles.

L'application de l'approche Bayesienne de la sélection de modèle conduit ici au critère d'évaluation d'un partitionnement bivarié supervisé, fourni dans la formule 3.

$$\log(N) + \log\left(C_{N+I_1-1}^{I_1-1}\right) + \log(N) + \log\left(C_{N+I_2-1}^{I_2-1}\right) + \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(C_{N_{i_1 i_2.}+J-1}^{J-1}\right) + \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(N_{i_1 i_2.}! / N_{i_1 i_2 1}! N_{i_1 i_2 2}! \ldots N_{i_1 i_2 J}!\right) \quad (3)$$

De façon similaire au cas de la discrétisation univariée, les deux premiers termes d'a priori correspondent au choix de la partition (nombre d'intervalles et bornes) de la première variable exogène. De même, les deux termes suivants correspondent au choix de la partition de la deuxième variable exogène. Le dernier terme d'a priori, en fin de première ligne, représente le choix de la distribution des valeurs endogènes dans chaque cellule. Le dernier terme de la formule 3, sur la deuxième ligne, représente la vraisemblance, c'est à dire la probabilité d'observer les valeurs de la variable endogène dans les cellules de la grille connaissant le modèle de partitionnement bivarié supervisé.

Le critère d'évaluation bivariée se généralise au cas des variables exogènes catégorielles, en remplaçant les termes de partition en intervalles (de type $\log(N) + \log(C_{N+I-1}^{I-1})$) par des termes de partitionnement en groupes de valeurs (de type $\log(V)+\log(B(V,I))$).

3.3 Algorithme d'optimisation

Nous proposons un algorithme d'optimisation, qui partant d'une solution initiale de partitionnement bivarié aléatoire, procède en alternant les optimisations partielles par variable:

1. initialiser une grille bivariée aléatoire, basée sur $O(N^{1/2})$ parties par variable,
2. tant que amélioration du critère, répéter:
   (a) optimiser la partition de X_1, la partition de X_2 étant fixée,

(b) optimiser la partition de X_2, la partition de X_1 étant fixée.

Dans le cas univarié, on remarque que les critères d'évaluation se décomposent de façon additive en un coût de partition et des coûts par partie. Le coût de partition $C^{(E)}$ ne dépend que des caractéristiques globales de l'ensemble d'apprentissage (nombre total d'individus, nombre total de valeurs des variables) et de la taille de la partition. Les coûts par partie $C^{(P)}{}_i$ ne dépendent que des caractéristiques locales d'une partie i (nombre d'individus de la partie par valeur endogène). Par exemple dans le cas de la discrétisation univariée, le critère d'évaluation de la formule 1 peut se réécrire au moyen d'un terme $C^{(E)}$ de coût de partition et d'une somme de termes $C^{(P)}{}_i$ de coûts par partie en posant

$$C^{(E)} = \log(N) + \log\left(C_{N+I-1}^{I-1}\right) \text{ et } C^{(P)}{}_i = \log\left(C_{N_{i.}+J-1}^{J-1}\right) + \log\left(N_{i.}!/N_{i1}!N_{i2}!...N_{iJ}!\right).$$

Les méthodes d'optimisation univariée exploitent l'additivité du critère, afin de passer d'une complexité en $O(N^3)$ à une complexité en $O(JN\log(N))$, le facteur J s'expliquant par le coût d'évaluation des termes pour chaque partie.

Prenons maintenant le cas de la discrétisation bivariée supervisée (formule 3) et montrons que le critère obtenu en fixant une des partitions se décompose de façon additive sur l'autre partition. En fixant par exemple la partition de la variable X_2, nous obtenons

$$C^{(E)} = \log(N) + \log\left(C_{N+I_1-1}^{I_1-1}\right) + \log(N) + \log\left(C_{N+I_2-1}^{I_2-1}\right) \text{ et}$$

$$C^{(P)}{}_i = \sum_{i_2=1}^{I_2} \log\left(C_{N_{i_1 i_2 .}+J-1}^{J-1}\right) + \sum_{i_2=1}^{I_2} \log\left(N_{i_1 i_2 .}!/N_{i_1 i_2 1}!N_{i_1 i_2 2}!...N_{i_1 i_2 J}!\right).$$

Le nombre de parties I_2 étant fixé (au même titre que le nombre de valeurs endogènes J), le coût de partition $C^{(E)}$ ne dépend effectivement que des caractéristiques globales de l'ensemble d'apprentissage et de la taille de la partition I_1. Les coûts par partie $C^{(P)}{}_i$ ne dépendent que des caractéristiques locales de chaque partie. Nous pouvons alors réutiliser l'algorithme de discrétisation univariée pour optimiser la partition de la variable X_1, la partition de X_2 étant fixée. La complexité algorithmique est en $O(JI_2N\log(N))$, le facteur I_2J s'expliquant par le coût d'évaluation des termes pour chaque partie.

Les expérimentations montrent que l'algorithme procédant par optimisations univariées alternées converge extrêmement rapidement, rarement en plus de deux itérations, ce qui confère à l'algorithme une complexité en $O(JN^{3/2}\log(N))$. Il est à noter que le choix initial de $O(N^{1/2})$ partie par variable est un choix heuristique, visant à assurer un bon compromis entre finesse de la partition initiale et complexité algorithmique.

4 Expérimentation

Cette section présente des résultats d'expérimentation, permettant d'évaluer l'impact de la méthode de partitionnement bivarié supervisé sur les performances en classification.

4.1 Protocole

Les expérimentations sont menées en utilisant 30 jeux de données de l'UCI (Blake et Merz, 1996) décrits en table 1, représentant une grande diversité de domaines, de nombres d'individus, de variables (numériques et/ou catégorielles) et de valeurs endogènes.

No.	Domaine	Indiv.	Var.	Val.	No.	Domaine	Indiv.	Var.	Val.
1	Abalone	4177	8	28	16	Letter	20000	16	26
2	Adult	48842	15	2	17	Mushroom	8416	22	2
3	Australian	690	14	2	18	PenDigits	10992	16	10
4	Breast	699	10	2	19	Pima	768	8	2
5	Crx	690	15	2	20	Satimage	6435	36	6
6	German	1000	24	2	21	Segmentation	2310	19	7
7	Glass	214	9	6	22	SickEuthyroid	3163	25	2
8	Heart	270	13	2	23	Sonar	208	60	2
9	Hepatitis	155	19	2	24	Spam	4307	57	2
10	HorseColic	368	27	2	25	Thyroid	7200	21	3
11	Hypothyroid	3163	25	2	26	TicTacToe	958	9	2
12	Ionosphere	351	34	2	27	Vehicle	846	18	4
13	Iris	150	4	3	28	Waveform	5000	21	3
14	Led	1000	7	10	29	Wine	178	13	3
15	Led17	10000	24	10	30	Yeast	1484	9	10

TAB. 1 – *Jeux de données de l'UCI, avec rappel du nombre d'individus, du nombre de variables exogènes et du nombre de valeurs endogènes.*

Afin d'évaluer la méthode d'analyse bivariée intrinsèquement, on introduit un nouveau type de classifieur appelé BestBivariate (B2). Ce classifieur recherche d'abord la meilleure paire de variable, celle qui maximise la probabilité que son modèle de partitionnement en grille explique la variable endogène. Pour classifier un individu en test, on recherche la cellule exogène associée aux valeurs de l'individu pour la paire de variables, et on prédit la valeur endogène majoritaire de la cellule (d'après les effectifs collectés en apprentissage). Si cette cellule était vide en apprentissage, on prédit la valeur endogène majoritaire de l'ensemble d'apprentissage. On évalue également le classifieur BestUnivariate (B1) qui procède de la même façon à partir de l'analyse univariée, et on rappelle pour référence le taux de bonne prédiction du classifieur majoritaire (M).

Afin d'évaluer l'impact de la méthode sur un classifieur multivarié, on évalue le classifieur Bayesien naïf (Langley et al, 1992), basé sur les prétraitements univariés (NB1) ou bivariés (NB2). On utilise également l'amélioration de ce classifieur (SNB1 et SNB2) décrite dans (Boullé, 2006b), qui incorpore d'une part une méthode régularisée de sélection de variables et d'autre part une agrégation de modèles, aboutissant à une pondération des variables[1]. L'exploitation de l'analyse bivariée est ici rudimentaire, en considérant chaque partitionnement bivarié comme une nouvelle variable calculée enrichissant l'espace de représentation.

Le taux de bonne prédiction en test est évalué au moyen d'une validation croisée stratifiée à 10 niveaux. Les différences significatives sont évaluées au seuil de 95% au moyen d'un test de Student.

4.2 Résultats

Les résultats d'évaluation sont résumés de façon synthétique dans le tableau 2, en reportant pour chaque méthode la moyenne arithmétique de son taux de bonne prédiction en test sur les 30 jeux de données de l'UCI. Le nombre de différences significatives vis-à-vis du classifieur SNB2 est également présenté, ainsi que le rang moyen de chaque méthode.

[1] Outil disponible en shareware sur http://www.francetelecom.com/en/group/rd/offer/software/technologies/middlewares/khiops.html.

	SNB2	NB2	SNB1	NB1	B2	B1	M
Moyenne	83.9%	81.9%	82.4%	81.4%	73.4%	67.6%	48.5%
Win/Draw/Loss		15/15/0	12/18/0	14/16/0	20/10/0	23/7/0	
Rang moyen	1.8	3.3	2.3	3.4	4.4	5.4	

TAB. 2 – *Valeur moyenne du taux de bonne prédiction en test, nombre de différences significatives du classifieur SNB2 comparé aux autres méthodes, et rang moyen des méthodes de classification sur les 30 jeux de données de l'UCI.*

On constate que le classifieur basé sur une seule variable (B1) est aussi performant que le meilleur classifieur multivarié (SNB2) dans environ un quart des cas (pas de différence significative dans 7 cas sur 30), et que celui basé sur deux variables uniquement (B2) atteint la meilleure performance obtenue dans un tiers des cas (10 cas sur 30). La figure 4 analyse plus finement les apports prédictifs du meilleur classifieur univarié (B1), du meilleur classifieur bivarié (B2) et du classifieur Bayesien naïf (NB1), avec comme référence le classifieur majoritaire (M). Le classifieur bivarié est systématiquement meilleur que le classifieur univarié, ce qui confirme la capacité de la méthode d'évaluation bivariée à correctement sélectionner une paire de variable performante. Il est toutefois dominé de façon significative par le classifieur Bayesien naïf, qui exploite l'ensemble de toutes les variables.

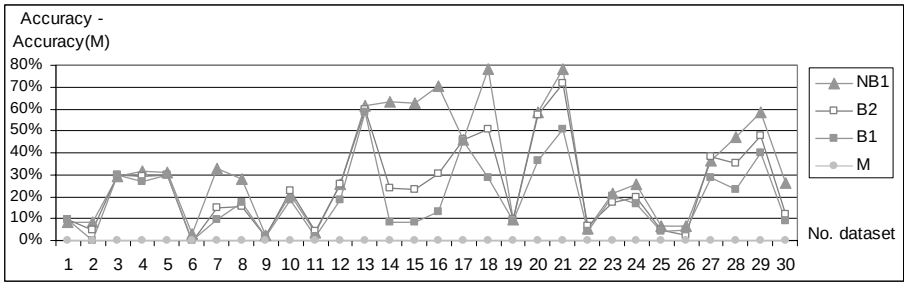


FIG. 4 – *Différence du taux de bonne prédiction par rapport au classifieur majoritaire (M) pour les classifieurs BestUnivariate (B1), BestBivariate (B2) et Bayesien naïf (NB1).*

La figure 5 analyse l'apport prédictif de la prise en compte de toutes les paires de variable (NB2) et de la sélection de variables (SNB1 et SNB2), en prenant pour référence le classifieur multivarié le plus simple (NB1). L'utilisation de paires de variables agrandit l'espace de représentation, ce qui permet potentiellement de détecter de nouvelles informations prédictives. En revanche, les informations redondantes déjà présentes en univarié sont ici multipliées, ce qui éloigne la représentation des données de l'hypothèse d'indépendance des variables sur laquelle repose le classifieur Bayesien naïf.

La figure 5 montre que ces deux phénomènes sont constatés sur les jeux de données de l'expérimentation quand on prend en compte les paires dans le classifieur NB2, avec de fortes dégradations de performances sur les jeux de données 1, 2, 6, 9, 22, 26, et de fortes améliorations sur les jeux de données 16, 18, 20, 27. La méthode de sélection de variables (Boullé, 2006b) utilisée dans SNB1 confirme son apport systématique, mais faible par rapport au classifieur Bayesien naïf NB1. Conjuguée à l'utilisation des paires de variables, le gain de

performance devient à la fois important, avec une amélioration moyenne de 2.5% (15% sur Letter), et significatif, avec 14 victoires significatives pour 0 défaites.

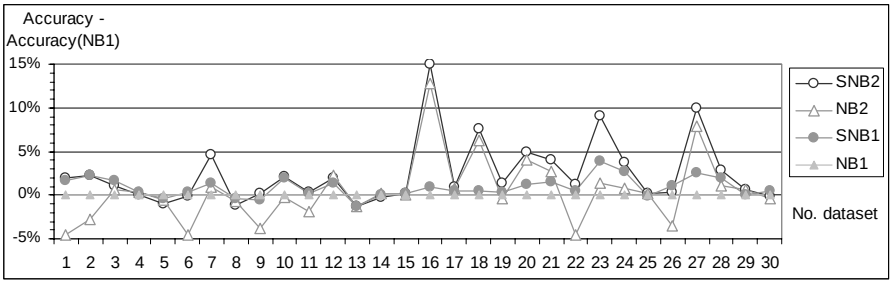


FIG. 5 – *Différence du taux de bonne prédiction par rapport au classifieur Bayesien naïf (NB1) pour le classifieur Bayesien naïf utilisant les paires de variables (NB2), et pour les classifieurs utilisant la sélection de variables (SNB1 et SNB2).*

5 Conclusion

La méthode d'évaluation bivariée supervisée présentée dans cet article se base sur un modèle de partitionnement (en intervalles ou groupes de valeurs) de chaque variable exogène, ce qui induit une partition bivariée. Cette partition bivariée permet de qualifier l'information apportée conjointement par les deux variables exogènes sur la variable endogène. Cette information est quantifiée au moyen d'une approche Bayesienne. Le critère d'évaluation obtenu est optimisé au moyen d'une heuristique gloutonne en alternant les améliorations partielles par variables.

Des évaluations intensives sur 30 jeux de données de l'UCI démontrent que le critère permet de sélectionner des paires de variables fortement informatives. L'ajout des paires de variables dans un classifieur Bayesien naïf n'est pas concluant en moyenne, les interactions constructives entre variables étant compensées par la redondance accrue de la représentation des données. En revanche, couplée avec une méthode performante de sélection de variables, la prise en compte des paires de variables apporte une amélioration systématique des performances, significative dans la moitié des cas.

De travaux futurs sont envisagés pour limiter le nombre de paires à évaluer et pour adapter les méthodes de classification à une prise en compte efficace des prétraitements bivariés.

Références

Bay, S.D. (2001). Multivariate Discretization for Set Mining. *Knowledge and Information System*, 3(4):491-512.

Blake, C.L. et C.J. Merz (1996). UCI Repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Boullé, M. (2005). A Bayes Optimal Approach for Partitioning the Values of Categorical Attributes. *Journal of Machine Learning Research*, 6:1431-1452.

Boullé, M. (2006a). MODL: a Bayes Optimal Discretization Method for Continuous Attributes, *Machine Learning*, 65(1):131-165.

Boullé, M. (2006b). Regularization and Averaging of the Selective Naïve Bayes Classifier. *2006 International Joint Conference on Neural Networks*, 2989-2997.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, et R. Wirth (2000). *CRISP-DM 1.0: step-by-step datamining guide*.

Govaert, G. et M. Nadif (2006). Classification d'un tableau de contingence et modèle probabiliste. *Revue des Nouvelles Technologies de l'Information*, 2:457-462,

Guyon, I. et A. Elisseeff (2003), An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157-1182.

Kohavi, R. et G. John (1997). Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273-324.

Langley, P., W. Iba et K. Thompson (1992). An analysis of Bayesian classifiers. *Proceedings of the 10th national conference on Artificial Intelligence*, AAAI Press, 223-228.

Muhlenbach, F. et R. Rakotomalala (2002). Multivariate Supervised Discretization, a Neighborhood Graph Approach. *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, 314-321.

Pyle, D (1999). *Data Preparation for Data Mining*. 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*, 39(154-155):81-98.

Saporta, G. (1990). *Probabilités analyse des données et statistique*. Editions TECHNIP.

Webb, G.I., J.R. Boughton et Z. Wang (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*, 58(1):5-24.

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

Zighed, D.A., G. Ritschard, W. Erray et V.M. Scuturici (2005). Decision trees with optimal joint partitioning. *International Journal of Intelligent System*, 20(7):693-718.

Summary

In the domain of data preparation for supervised learning, filter methods for variable selection are time efficient. However, their intrinsic univariate limitation prevents them from detecting redundancies or constructive interactions between variables. This paper introduces a new method to automatically, rapidly and reliably evaluate the predictive information of a pair of variables. It is based on a partitioning of each input variable, in intervals in the numerical case and in groups of values in the categorical case. The resulting input data grid allows to evaluate the correlation between the two input variables and the output variable. The best joint partitioning is searched owing to a Bayesian model selection approach. Intensive experiments demonstrate the benefits of the approach, especially the significant improvement of the classification accuracy.