Pouvoir explicatif et discriminant de variables et de tableaux de données symboliques

E. Diday

CEREMADE. Université Paris Dauphine

Résumé : Expliquer pour comprendre n'est pas discriminer pour apprendre. Plus précisément, on s'intéresse aux liens entre pouvoir explicatif d'une variable qualitative décrivant des classes par des diagrammes de fréquence et pouvoir discriminant de cette variable. On montre que la variable la plus explicative n'est pas nécessairement la plus discriminante et on énonce huit règles explicitant ces liens. On donne ensuite des critères permettant de sélectionner la variable à la fois la plus explicative et la plus discriminante. On donne aussi des critères permettant de mesurer le pouvoir explicatif et discriminant d'un tableau de données symboliques. Dans ces critères, on introduit l'entropie ou un critère de Gini afin d'accroître le pouvoir explicatif par des diagrammes de fréquence plus contrastés et concentrés. On utilise ces critères pour définir des indicateurs logiques ou numériques dont il faut optimiser les paramètres. On évoque enfin les champs potentiels d'applications industrielles.

1 Introduction

La statistique et plus précisément l'Analyse des données classiques (*Data Mining*), est fondée sur une description d'unités statistiques par des variables numériques et (ou) qualitatives. De façon plus générale, l'Analyse des données symboliques (ADS) considère comme unités statistique des classes dont les éléments (i.e. « individus ») ont des caractéristiques communes (des espèces de plantes ou d'animaux, des clients de même profil, des niveaux de dégradation d'ouvrages, des habitants d'une même région, etc.). Une même entité individuelle peut aussi varier (dans l'espace, le temps ou dans ses parties, ses différentes formes physiques, etc.) et être également considérée comme une classe formée de ses différentes variations. L'objectif premier d'une ADS est de décrire ces classes selon leurs multiples facettes. Cette description est obtenue en fusionnant les données initiales (qui décrivent les individus initiaux) sous forme d'un tableau non purement numérique (donc dit « symbolique ») formé de variables à valeurs intervalles, distributions, diagrammes de fréquence, histogrammes, suites de valeurs parfois pondérées, etc., exprimant la variabilité interne des classes. À ces variables issues de l'agrégation des données individuelles, on ajoute parfois des variables exprimant des propriétés propres aux classes.

Comme les données classiques sont un cas particulier de données symboliques, beaucoup de méthodes de l'ADS constituent une extension de méthodes classiques aux données symboliques. La théorie et la pratique de l'ADS ont été développées dans plusieurs ouvrages : Bock et Diday (2000), Billard et Diday (2006), Brito *et al.* (2007), Diday et Noirhomme (2008) ainsi que dans plusieurs articles de synthèse Billard et Diday (2003), Noirhomme et Brito (2012). Plusieurs essais de synthèse ont été proposés : confer, par exemple Billard et Diday (2003), Brito et Noirhomme (2015). Depuis 2012, quatre revues ont sorti des numéros spéciaux : *Statistical Analysis and Data Analysis* (Wiley, 2012), *RNTI* (Herman, 2013), *ADAC* (Springer, 2015), *IEEE Man and Cybernetic* (2016).
En dehors de l'extension des méthodes classiques aux données symboliques, des recherches et développement spécifiques à l'ADS sont nécessaires. C'est justement le cas du travail que nous présentons dans cet article où l'on s'intéresse au cas de tableaux de données symboliques dont les cases contiennent des diagrammes en bâtons.

On peut aboutir à un tel tableau soit à partir d'un tableau de données symboliques soit à partir des données initiales.
En partant d'un tableau de données symboliques, on peut transformer les variables symboliques qui ne sont pas à valeur diagramme en bâtons, en des variables symboliques de ce type. Ainsi, par exemple, une variable à valeur intervalle peut être transformée en un diagramme en bâtons à *k* modalités en associant à chaque intervalle réalisé, une loi uniforme et en découpant l'union des intervalles réalisés en *k* intervalles ; de même une liste peut être transformée en un diagramme en bâtons en considérant les termes de la liste comme équiprobables.
Si l'on part maintenant, d'un tableau de données classiques défini par des variables quantitatives et / ou qualitatives, on peut transformer les variables quantitatives en variables qualitatives. De façon à ce que les diagrammes obtenus sur les classes soient bien discriminés, il faut utiliser des algorithmes permettant de réaliser au mieux la discrétisation : voir, par exemple, Fisher W. (1958), Fayyad et Irani (1993). Dans Diday *et al.* (2013), cette discrétisation se fait de façon que les diagrammes en bâtons obtenus sur les classes données les discriminent au mieux.

Pour décrire des classes par des variables qualitatives, on aboutit à un tableau de données dites « symboliques » car les cases d'un tel tableau ne contiennent pas des nombres ou des catégories mais des diagrammes de fréquence. Les variables de ce tableau sont donc à valeur diagrammes de fréquences. Plus l'écart (au sens de la distance L_1, par exemple) entre ces diagrammes de fréquences est grand plus la variable est dite « explicative ». À l'inverse, on peut aussi décrire une variable qualitative par la variable classe. Chacune de ses modalités peut alors être décrite par le diagramme de fréquences induit par les classes. Plus l'écart entre ces diagrammes de fréquence est grand plus la variable est dite discriminante. Nous montrons d'abord que la variable la plus explicative n'est pas nécessairement la plus discriminante et plus généralement nous énonçons huit règles liant le pouvoir explicatif et discriminant d'une variable. Nous illustrons ces résultats par deux exemples pratiques. Si les diagrammes de fréquences sont contrastés et concentrés, on obtient un plus grand pouvoir explicatif qui s'exprime à l'aide d'un critère de Gini ou de l'entropie. On en déduit des critères permettant de sélectionner les meilleures variables symboliques ou mesurant le pouvoir explicatif d'un tableau de données symboliques en tenant compte à la fois du pouvoir explicatif et discriminant des variables symboliques. On propose ensuite des indicateurs logiques et numériques à la fois explicatifs et discriminants. Finalement, on indique des pistes pour des applications éventuelles.

2 Les données

On dispose au départ de variables qualitatives notées *C, Y, Z* définies sur une population de *n* individus (voir la table 1). L'objectif de cet article est d'indiquer dans quelles conditions on peut dire que si *Y* décrit les classes de *C* (voir la table 2) de la façon la plus « explicative » (en un sens qui sera précisé), alors *Y* (voir la table 3) est aussi celle dont les classes sont les plus « discriminantes » (en un sens qui sera précisé), de la variable *C*.

Individus	***C***	***Y***	***Z***
Ind $_i$	Cj	Y_l	Z_t
Ind n			

Table 1 : tableau de données initiales où *C, Y, Z* sont des variables qualitatives à respectivement *k, k', k''* modalités.

À chaque modalité d'une variable qualitative, on peut associer une classe d'individus, ce qui permet de décrire la variable qualitative *C* par les variables symboliques Y_C et Z_C à valeur

diagramme de fréquence des modalités $Y_1, \ldots, Y_{k'}$ de Y et $Z_1, \ldots, Z_{k''}$ de Z dans chaque classe induite par chaque modalité de C.

Ainsi, la description (appelée parfois « objet symbolique ») notée dCi/Y de la classe Ci par un diagramme en bâtons s'écrit :

$dCi/Y = (Pr(Y1 \cap Ci)/Pr(Ci), \ldots, Pr(Yk' \cap Ci)/Pr(Ci))$ où $Pr(Ym \cap Ci)$ est le rapport du nombre d'individus de la classe Ci ayant pris la modalité Ym divisé par le nombre total d'individus. $Pr(Ci)$ est le cardinal de Ci également divisé par le nombre total d'individus. Cela revient à dire que $Pr(Yl \cap Ci)/Pr(Ci)$ est la fréquence de la modalité Ym dans la classe Ci. Plus généralement, dans le cas de plusieurs variables descriptives, chaque classe, considérée comme un « objet », est décrite par un « objet symbolique » défini par un vecteur de diagrammes en bâtons.

C	Y_C	Z_C
C_1	$Pr(Y_1 \cap C_1)/Pr(C_1), \ldots, Pr(Y_{k'} \cap C_1)/Pr(C_1)$	$Pr(Z_1 \cap C_1)/Pr(C_1), \ldots, Pr(Z_{k''} \cap C_1)/Pr(C_1)$
C_k	$Pr(Y_1 \cap C_k)/Pr(C_k), \ldots, Pr(Y_{k'} \cap C_k)/Pr(C_k)$	$Pr(Z_1 \cap C_k)/Pr(C_k), \ldots, Pr(Z_{k''} \cap C_k)/Pr(C_k)$

Table 2. Tableau *C/V* permettant de calculer le pouvoir explicatif : les k classes de C sont décrites par les variables symboliques Y_C et Z_C prises parmi un ensemble V de variables symboliques.

On peut de même décrire la variable Y par la variable symbolique C_Y à valeur diagramme de fréquence des modalités $C_1, \ldots, C_k$ de C dans chaque classe induite par chaque modalité de Y. On peut faire de même avec la variable Z, ce qui produit la table 3.

Y	C_Y
Y_1	$Pr(C_1 \cap Y_1)/Pr(Y_1), \ldots, Pr(C_k \cap Y_1)/Pr(Y_1)$
$Y_{k'}$	$Pr(C_1 \cap Y_{k'})/Pr(Y_{k'}), \ldots, Pr(C_k \cap Y_{k'})/Pr(Y_{k'})$

Z	C_Z
Z_1	$Pr(C_1 \cap Z_1)/Pr(Z_1), \ldots, Pr(C_k \cap Z_1)/Pr(Z_1)$
$Z_{k'}$	$Pr(C_1 \cap Z_{k'})/Pr(Z_{k'}), \ldots, Pr(C_k \cap Z_{k'})/Pr(Z_{k'})$

Table 3. Tableau *Y/C* et *Z/C* permettant de calculer le pouvoir discriminant : les k' (resp. k'') modalités de la variable Y (resp. Z) sont décrites par les variables symboliques C_Y et C_Z à valeur diagramme de fréquence.

3 Lien entre pouvoir explicatif et pouvoir discriminant.

3.1 Les huit règles liant pouvoir explicatif et discriminant

Nous voulons étudier le lien entre pouvoir explicatif des classes données et le pouvoir décisionnel sur ces classes par des variables qualitatives. Nous disons qu'une variable Y explique bien des classes si elle les décrit de façon bien distinctive. Nous disons qu'une variable Y discrimine bien des classes, si la variable « classe » associée à ces classes décrit les classes de la variable Y de façon bien distinctive. Finalement, on se pose la question suivante : dans quelles conditions la variable la plus explicative de deux classes est aussi la plus discriminante ?

Pour commencer, on veut comparer le pouvoir discriminant et le pouvoir explicatif de deux variables qualitatives Y et Z en utilisant la distance L_1. On suppose que les deux variables sont chacune à deux modalités et donc que : $k=k'=k=2$.
On pose : $a_y = Pr(Y_1 \cap C_1)$, $b_y = Pr(Y_1 \cap C_2)$, $d_y = Pr(Y_2 \cap C_1)$, $c_i = Pr(C_i)$, $y_i = Pr(Y_i)$.
Remarquons que : $c_1 = d_y + a_y = d_z + a_z$, $y_1 = a_y + b_y$ et que $y_1 + y_2 = c_1 + c_2 = c$.
On considère que $D_{C/Y}$ exprime le pouvoir explicatif de Y et que $D_{Y/C}$ exprime le pouvoir discriminant de Y pour la variable classe C.

Avec la distance L_1, on a ainsi pour le pouvoir explicatif de Y :

$$D_{C/Y}(C_1, C_2) = (\Sigma_{j=1,2} |Pr(C_1 \cap Y_j) / Pr(C_1) - Pr(C_2 \cap Y_j) / Pr(C_2) |)/2$$

et pour le pouvoir discriminant de Y :

$$D_{Y/C}(Y_1, Y_2) = (\Sigma_{j=1,2} |Pr(C_j \cap Y_1) / Pr(Y_1) - Pr(C_j \cap Y_2) / Pr(Y_2)|)/2$$

Notons que $D_{Y/C}(Y_1, Y_2)$ varie entre 0 et 1.

Considérons les tables suivantes :

C/Y	***Y***	
Classes	Y_1	Y_2
C_1	a_y / c_1	$1- a_y/c_1$
C_2	b_y / c_2	$1- b_y / c_2$

$Tab_{C/Y}$ description de C par Y

C/Z	***Z***	
Classes	Z_1	Z_2
C_1	a_z/c_1	$(c_1- a_z)/c_1$
C_2	b_z / c_2	$1- b_z / c_2$

$Tab_{C/Z}$ description de C par Z

Figure 1 : description des classes de la variable C par les variables symboliques Y et Z à valeur diagramme.

Y/C	***C***	
Catégories	C_1	C_2
Y_1	a_y/y_1	$1- a_y/y_1$
Y_2	d_y/y_2	$1-d_y/y_2$

$Tab_{Y/C}$: description de Y par C

Z/C	***C***	
Catégories	C_1	C_2
Z_1	a_z / z_1	$1- a_z/z_1$
Z_2	d_y/ z_2	$1-d_y/ z_2$

$Tab_{Z/C}$: description de Z par C

Figure 2 : description des classes des variables Y et Z par la variable symbolique C à valeur diagramme.

La question que l'on se pose maintenant est la suivante : entre Y et Z peut-on dire que la variable la plus explicative soit également la plus discriminante ?

La réponse à cette question est importante car dans le cas où les deux pouvoirs sont très liés, l'analyse peut être faite uniquement sur la table qui permet de calculer l'un des deux pouvoirs. Par exemple, il suffit de sélectionner les variables de meilleur pouvoir explicatif sans se préoccuper des variables de meilleur pouvoir discriminant puisque dans ce cas ce seront les mêmes.

Remarquons que, dans le cas où les données ont été discrétisées, les résultats seront dépendants de la discrétisation, néanmoins pour une discrétisation jugée bonne par l'utilisateur, les résultats seront intéressants et utiles. En utilisant la distance L_1, nous allons donner des conditions suffisantes pour que la variable de plus grand pouvoir explicatif soit ou ne soit pas celle de plus grand pouvoir discriminant des classes de C. Avant d'énoncer la proposition, nous allons énoncer trois lemmes.

Lemme 1

On a les deux propriétés suivantes :

$$D_{C/Y} / D_{C/Z} = |a_y c_2 - b_y c_1| / |a_z c_2 - b_z c_1|$$

et

$$D_{C/Y}/\ D_{C/Z} = |a_y c - y_1 c_1| \,/\, |a_z c - b_z c_1|.$$

Démonstration

On a $D_{C/Z} = (|a_z /c_1 - b_z/c_2| + |(1 - a_z/c_1 - (1 - b_z/c_2|)/2 = |a_z /c_1 - b_z/c_2| = |a_z c_2 - b_z c_1|/c_1c_2$

De même $D_{C/Y} = |a_y/c_1 - b_y/c_2| = |a_y c_2 - b_y c_1|/c_1c_2$

Donc $D_{C/Y}/\ D_{C/Z} = |a_y c_2 - b_y c_1| \,/\, |a_z c_2 - b_z c_1|$

Comme $a_y c_2 - b_y c_1 = a_y (c - c_1) - b_y c_1 = a_y c - c_1 (a_y + b_y) = a_y C - c_1 y_1$

On a : $D_{C/Y}/\ D_{C/Z} = |\ a_y c - y_1 c_1| \,/\, |\ a_z c - z_1 c_1|$

Cqfd

Lemme 2

On a les deux propriétés suivantes :

$$D_{Y/C}/\ D_{Z/C} = |a_y c_2 - b_y c_1|\ z_1\ z_2\ \ /\ |a_z c_2 - b_z c_1|\ y_1 y_2$$

et

$$D_{Y/C}/\ D_{Z/C} = |a_y C - y_1 c_1|\ z_1 z_2\ \ /\ |a_z C - z_1 c_1|\ \ y_1 y_2$$

Démonstration

On a $D_{Z/C} = (|a_z /z_1 - d_z/z_2| + |(1 - a_z /z_1 - (1 - d_z /z_2))/2 = |a_z /z_1 - d_z/z_2|$

or $|a_z /z_1 - d_z/z_2| = |a_z\ z_2 - d_z z_1|/\ z_1\ z_2$

comme $z_1 = (a_z + b_z),\ d_z = (c_1 - a_z)$ et $z_2 = \ (c_1 + c_2 - a_z - b_z)|$

on a : $|a_z /z_1 - d_z/z_2| = \ \ |a_z\ (c_1 + c_2 - a_z - b_z) - (c_1 - a_z)\ (a_z + b_z) \,/\, z_1\ z_2$

d'où, finalement : $D_{Z/C} = |a_z c_2 - b_z c_1| \,/\, z_1\ z_2$

De même $D_{Y/C} = |a_y c_2 - b_y c_1|/\ y_1 y_2$

d'où : $D_{Y/C}/\ D_{Z/C} = |a_y c_2 - b_y c_1|\ z_1\ z_2\ \ /\ |a_z c_2 - b_z c_1|\ y_1 y_2$

et donc : $D_{Y/C}/\ D_{Z/C} = |a_y C - y_1 c_1|\ z_1 z_2\ \ /\ |a_z C - z_1 c_1|\ \ y_1 y_2.$

Cqfd

Comme conséquence de ces deux lemmes, on a la proposition suivante :

Lemme 3

$$D_{C/Y} / D_{C/Z} = (y_1\ y_2\ /\ z_1\ z_2)\ (D_{Y/C} / (D_{Z/C}) \qquad \textbf{(1)}$$

$$D_{Y/C} / D_{Z/C} = (z_1\ z_2 / y_1\ y_2)\ (D_{C/Y}) / D_{C/Z}) \qquad \textbf{(2)}$$

Démonstration

$D_{C/Y}/\ D_{C/Z} = |a_y c_2 - b_y c_1| \,/\, |a_z\ c_2 - b_z c_1|$ d'après le lemme 1

et

$D_{Y/C}/\ D_{Z/C} = |a_y c_2 - b_y c_1|\ z_1\ z_2\ \ /\ |a_z c_2 - b_z c_1|\ y_1 y_2$ d'après le lemme 2, on a donc

$$D_{C/Y} / D_{C/Z} = (D_{Y/C}\, y_1\, y_2) / (D_{Z/C}\, z_1\, z_2) \text{ ou } D_{Y/C} / D_{Z/C} = (z_1\, z_2 / y_1 y_2)\, (D_{C/Y}) / D_{C/Z}).$$

Cqfd

Remarque : de ce résultat, nous pouvons déduire d'après la relation (2) que le pouvoir explicatif et le pouvoir discriminant d'une variable symbolique *Y* sont proportionnels. Il résulte aussi de la relation (2) que si l'on a $D_{C/Y} / D_{C/Z} < 1$ et $z_1z_2 / y_1y_2 < 1$, alors nécessairement on a : $D_{Y/C} / D_{Z/C} < 1$. Autrement dit, $D_{C/Y} < D_{C/Z}$ avec $z_1z_2 < y_1y_2$ implique $D_{Y/C} < D_{Z/C.}$ et donc la variable la plus explicative (*Z*) est aussi la plus discriminante. Par contre, du fait de la relation (1), si $D_{C/Y} / D_{C/Z} > 1$ et $y_1y_2 / z_1z_2 > 1$ alors l'inégalité $D_{Y/C} < D_{Z/C.}$ n'est pas certaine, on peut seulement dire qu'elle est possible.

La proposition suivante généralise ces cas à tous les cas possibles.

Proposition 1

On a les 8 règles suivantes :

Si \ alors	$D_{Z/C} < D_{Y/C}$	$D_{Z/C} > D_{Y/C}$
$D_{C/Z} < D_{C/Y}$ et $y_1y_2 \leq z_1z_2$	vrai	faux
$D_{C/Z} < D_{C/Y}$ et $y_1y_2 > z_1z_2$	possible	possible
$D_{C/Z} > D_{C/Y}$ et $y_1y_2 < z_1z_2$	possible	possible
$D_{C/Z} > D_{C/Y}$ et $y_1y_2 \Sigma z_1z_2$	vrai	faux

Figure 3 : la première case « vrai » de la table exprime « une condition suffisante pour que la meilleure variable explicative soit aussi la meilleure variable discriminante des classes au sens de la distance de la valeur absolue est que : $y_1y_2 < z_1z_2$ ».

Démonstration

Il suffit de poser $u = D_{C/Y} / D_{C/Z}$, $v = D_{Y/C} / D_{Z/C}$, $w = y_1y_2/z_1z_2$. On a alors $u = vw$.

Si $u>1$ et $w<1$, on a forcément $v>1$. Ce qui prouve qu'une condition suffisante pour que la meilleure variable explicative soit aussi la meilleure variable discriminante des classes au sens de la distance de la valeur absolue est que : $y_1y_2 > z_1z_2$. Il est impossible d'avoir $v<1$ car alors $u>1$ serait le produit de deux nombres inférieurs à 1. Par contre, si $u>1$ et $w>1$, il est possible d'avoir *v* supérieur ou inférieur à *1*. Les quatre derniers cas des deux lignes qui suivent, se démontrent de façon analogue.

Cqfd

Remarque pratique : pour choisir la meilleure variable, on peut toujours se placer dans le cas où $y_1y_2 < z_1z_2$, on sait alors que $D_{C/Z} < D_{C/Y}$ implique $D_{Z/C} < D_{Y/C}$, on est alors certain que *Y* est la meilleure variable aussi bien pour l'explication que pour la discrimination des classes. Si $D_{C/Z} > D_{C/Y}$, on ne peut pas se prononcer.

3.2 Exemples illustratifs montrant différents liens entre le pouvoir explicatif d'une variable et son pouvoir décisionnel

Exemple 1

Considérons le tableau de données suivant :

Individus	***Y***		***Z***		***C***	
	Y1	***Y2***	***Z1***	***Z2***	***C1***	***C2***
I1	*1*	*0*	*0*	*1*	*1*	*0*
I2	*1*	*0*	*1*	*0*	*1*	*0*
I3	*1*	*0*	*1*	*0*	*0*	*1*
I4	*1*	*0*	*0*	*1*	*0*	*1*
I5	*0*	*1*	*0*	*1*	*0*	*1*
I6	*0*	*1*	*0*	*1*	*0*	*1*
I7	*0*	*1*	*0*	*1*	*0*	*1*
I8	*0*	*1*	*0*	*1*	*1*	*0*
I9	*0*	*1*	*0*	*1*	*1*	*0*
I10	*0*	*1*	*0*	*1*	*1*	*0*
I11	*0*	*1*	*0*	*1*	*1*	*0*
I12	*0*	*1*	*0*	*1*	*1*	*0*
I13	*0*	*1*	*0*	*1*	*1*	*0*
I14	*0*	*1*	*0*	*1*	*1*	*0*
I15	*0*	*1*	*0*	*1*	*1*	*0*
I16	*0*	*1*	*0*	*1*	*1*	*0*
I17	*0*	*1*	*0*	*1*	*1*	*0*
I18	*0*	*1*	*0*	*1*	*1*	*0*
I19	*0*	*1*	*0*	*1*	*1*	*0*
I20	*0*	*1*	*0*	*1*	*1*	*0*
Total	$y_1 = 4$	$y_2 = 16$	$z_1 = 2$	$z_2 = 18$	*15*	*5*
Paramètres	$a_y = 1/2$ $b_y = 1/2$	$d_y = 13/27$	$a_z = 1/2$ $b_z = 1/2$	$d_z = 14/27$	*c1= 15/20*	*c2= 5/20*

Table 4 : exemple de table où la variable la plus explicative (*Y*) n'est pas la variable la plus prédictive.

Selon les données de la table 4, on a : $y_1y_2 / z_1z_2 = 4x16/2x18 = 16/9 > 1$ donc d'après la proposition précédente (case 1 de la table 3) : $D_{C/Z} < D_{C/Y}$ doit impliquer $D_{Z/C} < D_{Y/C}$. Vérifions-le sur cet exemple.

Pour cela il suffit de vérifier que l'on a simultanément $D_{C/Z} < D_{C/Y}$ et $D_{Z/C} < D_{Y/C}$.

classes	Y_1	Y_2
C_1	$a_y / c_1 = 2/15$	$1- a_y / c_1 = 13/15$
C_2	$b_y / c_2 = 2/5$	$1- b_y / c_1 = 3/5$

classes	Z_1	Z_2
C_1	$a_z / c_1 = 1/15$	$1- a_z / c_1 = 14/15$
C_2	$b_z / c_2 = 1/5$	$1- b_z / c_1 = 4/5$

Figure 4a : résultats issus de la table 1 pour les tables : $Tab_{C/Y}$ et $Tab_{C/Z}$ avec les données de la table 1.

$D_{C/Z} = |1/15-1/5|+|14/15-4/5| = |5-15|/75 + |56-60|/75 = 10/75$

$< D_{C/Y} = |2/15-2/5|+|13/15-3/5| = (|10-30| + |39-45|)/75 = 26/75$

Y	***C₁***	***C₂***
Y_1	$a_y /y_1= 2/4$	$(1- a_y) /y_1= 1/2$
Y_2	$d_y /y_2= 13/15$	$(1- d_y) /y_2=2/15$

Z	***C₁***	***C₂***
Z_1	$a_z /z_1= 1/2$	$(1- a_z) /z_1= 1/2$
Z_2	$d_z /z_2= 14/15$	$(1- d_z) /z_2=1/15$

Figure 4b : résultats issus de la table 4 pour les tables $Tab_{Y/C}$ et $Tab_{Z/C}$.

$D_{Y/C}= |1/2-13/15|+|1/2-2/15| = |30-26|/30+|15-4|/30 = 15/30 = ½$

$D_{Z/C}= |1/2-14/15|+|1/2-1/15| = |15-28|/30+|15-30|/30 = 23/30$

On a donc bien $y_1y_2 / z_1z_2= 16/9 > 1$ et $D_{C/Z} < D_{C/Y}$ qui implique $D_{Z/C} < D_{Y/C}$.

Exemple 2

On a $c= 100$, $c_1 = 80$, $c_2 = 20$,

Pour la variable *Y*, on a : $a_y = 9$, $b_y =3$, $y_1 = 12$, $y_2 = 88$.

Pour la variable *Z*, on a : $a_z = 39$, $b_z =11$, $z_1 = 50$, $z_2 = 50$.

$D_{C/Y}/ D_{C/Z} = |a_yc_2- b_yc_1| / |a_z c_2- b_zc_1| = |9x20- 3x80|/ |39x20-11x80 | = 60/100= 3/5<1$

$D_{Y/C} / D_{Z/C} = (z_1 z_2 / y_1 y_2)\ D_{C/Y} / D_{C/Z} = (2500/ 1056) x 3/5 = 7500/5280= 125/88 >1.$

Il en résulte donc que : $D_{C/Y}(C_1, C_2) < D_{C/Z}(C_1, C_2)$ et $D_{Y/C}(Y_1, Y_2)> D_{Z/C}(Z_1, Z_2)$

Donc, pour cet exemple aussi, la variable la plus explicative n'est pas la variable la plus discriminante.

Exemple 3

Dans cet exemple, la variable la plus explicative est la variable la plus discriminante**.** Vérifions que la table 5 suivante correspond à ce cas :

i	***Y***	***Z***	***C***
i_1	*1*	*0*	*1*
i_2	*1*	*0*	*1*
i_3	*1*	*1*	*0*
i_4	*0*	*1*	*1*
i_5	*0*	*0*	*1*
i_6	*0*	*0*	*0*
i_7	*0*	*0*	*0*

Table 5

classes	**Y_1**	**Y_2**
C_1	$a_y /c_1 =$ 2/4	$1- a_y /c_1 =1/2$
C_2	$b_y /c_2 =$ 1/3	$1- b_y /c_1 = 2/3$

classes	**Z_1**	**Z_2**
C_1	$a_z /c_1 = 1/4$	$1- a_z /c_1 = 3/4$
C_2	$b_z /c_2 = 1/3$	$1- b_z /c_1 = 2/3$

Figure 5 : $TAB_{C/Y}$ et $TAB_{C/Z}$ issues de la table 4.

Y	**C_1**	**C_2**
Y_1	$a_y / y_1 = 2/3$	$(1- a_y) / y_1 = 1/3$
Y_2	$d_y / y_2 = 2/4$	$(1- d_y) / y_2 = 1/2$

Z	**C_1**	**C_2**
Z_1	$a_z / z_1 = 1/2$	$(1- a_z) / z_1 = 1/2$
Z_2	$d_z / z_2 = 3/5$	$(1- d_z) / z_2 = 2/5$

Figure 6 : $TAB_{Y/C}$ et $TAB_{Z/C}$ issues de la table 4.

$D_{C/Y} = (|1/2-1/3|+|1/2-2/3|)/2 = 1/6$, $D_{C/Z} = (|1/4-1/3|+|3/4-2/3|)/2 = 1/12$

donc $D_{C/Y} / D_{C/Z} = 2$

$D_{Y/C} = (|2/3-1/2|+|1/3-1/2|)/2 = 1/6$, $D_{Z/C} = (|1/2-3/5|+|1/2-2/5|)/2 = 1/10$

donc $D_{Y/C} / D_{Z/C} = 10/6 = 5/3$

On a $C = 7$, $c_1 = 4$, $c_2 = 3$.

Pour la variable Y, on a : $a_y = 2$, $b_y = 1$, $y_1 = 3$, $y_2 = 4$,

Pour la variable Z, on a : $a_z = 1$, $b_z = 1$, $z_1 = 2$, $z_2 = 5$,

D'après le lemme 1, on a : $D_{C/Y} / D_{C/Z} = |a_y c_2 - b_y c_1| / |a_z c_2 - b_z c_1|$

Comme $(a_y c_2 - b_y c_1) = (6-4) = 2$ et $|a_z c_2 - b_z c_1| = 1$, il en résulte que :

$D_{C/Y} / D_{C/Z} = 2$ donc $D_{C/Z} < D_{C/Y}$.

D'après le lemme 2 et la formule (2) on a :

$D_{Y/C} / D_{Z/C} = (z_1 z_2 D_{C/Y}) / (y_1 y_2 D_{C/Z}) = 2x10/12 = 5/3$ donc $D_{Z/C} < D_{Y/C}$.

On se trouve ainsi dans le cas de la ligne 2, première colonne de la table de la figure 3 où :

$D_{C/Z} < D_{C/Y}$ *et* $D_{Z/C} < D_{Y/C}$ avec $y_1 y_2 = 12 > 10 = z_1 z_2$.

Autrement dit, la variable la plus explicative (Y) est aussi la variable la plus discriminante dans ce cas.

4 Critère de sélection de variables symboliques à la fois explicatives et discriminantes

4.1 Sélection dans le cas de variables binaires basée sur la comparaison avec les autres variables.

On peut définir un critère de sélection à maximiser qui est d'autant plus grand que le pouvoir explicatif et décisionnel d'une variable Y est grand par rapport à celui des autres variables Z. Ce critère noté « Sel_1 » s'écrit de la façon suivante :

$Sel_1 (Y') = \Sigma_{Z \in V} (D_{Y'/C} / D_{Z/C}) \times (D_{C/Y'} / D_{C/Z})$ où V est l'ensemble des variables sélectionnables autres que Y'.

Proposition 2

$$\boldsymbol{Sel_1 (Y') = \Sigma_{Z \in V} (z_1 z_2 / y'_1 y'_2) \ (D_{C/Y'} / D_{C/Z})^2}$$

Démonstration

D'après le lemme 3, on a :

$D_{Y'/C} / D_{Z/C} = (z_1 z_2 / y'_1 y'_2)\ D_{C/Y'} / D_{C/Z}$ dont on peut déduire :

$Sel_1 (Y') = \Sigma_{Z \in V} (z_1 z_2 / y'_1 y'_2)\ D_{C/Y'} / D_{C/Z}\ x\, D_{C/Y'} / D_{C/Z})$, soit finalement :

$Sel_1 (Y') = \Sigma_{Z \in V} (z_1 z_2 / y'_1 y'_2)\ (D_{C/Y'} / D_{C/Z})^2$

Cqfd

La meilleure variable est alors la variable Y qui maximise $Sel_1 (Y')$.

4.2 Sélection de variables binaires avec prise en compte du contraste avec le critère de Gini.

On dit qu'un diagramme est « contrasté » si ses fréquences sont soit grandes, soit petites. Quand la dissimilarité n'est pas maximum, on peut se trouver devant différents cas de contraste. Si l'on veut améliorer le pouvoir explicatif des diagrammes de fréquence en augmentant ce contraste, aussi bien au niveau des tableaux descriptifs que des tableaux discriminants des classes de la variable C, on peut utiliser un critère de Gini. Avec ce critère, la qualité descriptive d'une variable Y s'écrit :

Posons $D'_{C/Y} (C_1, C_2) = D_{C/Y} (C_1, C_2) / U_y$

avec $U_y = Pr(Y_1 \cap C_1) Pr(Y_2 \cap C_1) / (Pr(C_1))^2 + Pr(Y_1 \cap C_2) Pr(Y_2 \cap C_2) / (Pr(C_2))^2$

d'où : $U_y = (a_y(c_1 - a_y)\, c^2_2 + (y_1 - a_y)(c_2 - (y_1 - a_y))\, c^2_1\ ) / c^2_1\, c^2_2$

On pose $u_y = a_y(c_1 - a_y)\, c^2_2 + (y_1 - a_y)(c_2 - (y_1 - a_y))\, c^2_1$

d'où $D'_{C/Y} (C_1, C_2) = c^2_1\, c^2_2\, D_{C/Y} / u_y$

et de même, en remplaçant partout y par z dans u_y, on obtient :

$U_z = Pr(Z_1 \cap C_1) Pr(Z_2 \cap C_1) + Pr(Z_1 \cap C_2) Pr(Z_2 \cap C_2)$

et $D'_{C/Z}(C_1, C_2) = c^2_1\, c^2_2\, D_{C/Z} / u_z$

d'où : $\boldsymbol{D'_{C/Y} / D'_{C/Z} = (u_z / u_y)\ D_{C/Y} / D_{C/Z}}$ **(3)**

Posons maintenant : $D'_{Y/C} (Y_1, Y_2) = D'_{Y/C} (Y_1, Y_2) / E_y$ avec

$E_y = Pr(Y_1 \cap C_1) Pr(Y_1 \cap C_2) + Pr(Y_2 \cap C_1) Pr(Y_2 \cap C_2)$,

$E_y = a_y (y_1 - a_y) / y^2_1 + (c_1 - a_y)(y_2 - (c_1 - a_y)) / y^2_2$ ou

$E_y = ((a_y (y_1 - a_y)\, y^2_2 + (c_1 - a_y)(y_2 - (c_1 - a_y)\, y^2_1)) / y^2_1\, y^2_2$

Posons $e_y = (a_y (y_1 - a_y)\, y^2_2 + (c_1 - a_y)(y_2 - (c_1 - a_y))\, y^2_1)$ d'où : $E_y = e_y / y^2_1\, y^2_2$

et, de même : $E_z = Pr(Z_1 \cap C_1) Pr(Z_1 \cap C_2) + Pr(Z_2 \cap C_1) Pr(Z_2 \cap C_2)$,

$E_z = e_z / z^2_1\, z^2_2$

d'où : $D'_{Y/C} (Y_1, Y_2) = c^2_1\, c^2_2\, D_{C/Y} / u_y$

et de même $D'_{Z/C}(Z_1, Z_2) = c^2_1\, c^2_2\, D_{C/Z} / u_z$

d'où :

$D'_{Y/C}/ D'_{Z/C} = (E_Z / E_Y) D_{Y/C}/ D_{Z/C}$

$D'_{Y/C} = y^2_1 y^2_2 D_{Y/C} / e_Y$ et $D'_{Z/C} = z^2_1 z^2_2 D_{Z/C} / e_z$

donc, d'après (2) : $D'_{Y/C}/ D'_{Z/C} = (y^2_1 y^2_2 e_z / z^2_1 z^2_2 e_y) (z_1 z_2 /y_1 y_2) (D_{C/Y})/ D_{C/Z})$

$D'_{Y/C}/ D'_{Z/C} = y^2_1 y^2_2 e_z D_{Y/C} / z^2_1 z^2_2 e_y D_{Z/C}$

et, finalement : $\boldsymbol{D'_{Y/C}/ D'_{Z/C} = (y_1 y_2 e_z / z_1 z_2 e_y) (D_{C/Y} / D_{C/Z})}$ **(4).**

Nous pouvons maintenant définir un critère exprimant à la fois le pouvoir explicatif et le pouvoir discriminant de *Y* par rapport à *Z* comme pour le critère Sel_1 mais avec un pouvoir explicatif encore plus grand prenant cette fois en compte le « contraste » des diagrammes de fréquence à l'aide de l'indice de Gini. Posons :

$Sel_2 (Y') = \Sigma_{Z \in V} (D'_{Y'/C} / D'_{Z/C} \; x D'_{C/Y'}/ D'_{C/Z})$, donc d'après (3) et (4)

$= \Sigma_{Z \in V} ((y'_1 y'_2 e_z / z_1 z_2 e_{y'}) (D_{C/Y'}/ D_{C/Z}) (u_z /u_y) D_{C/Y'}/ D_{C/Z})$

Finalement :

$$\boldsymbol{Sel_2 (Y') = \Sigma_{Z \in V} (y'_1 y'_2 u_z e_z / z_1 z_2 \; u_{y'} e_{y'}) (D_{C/Y'} / D_{C/Z})^2} \qquad \textbf{(5)}$$

Pour éviter les problèmes de division par 0, on peut ajouter 1 au dénominateur (ce qui ne modifie pas la plage de variation entre 0 et 1) pour obtenir finalement :

$$\boldsymbol{Sel_3 (Y') = (\Sigma_{Z \in V} (y'_1 y'_2 u_z e_z D^2_{C/Y'}/ (1+ z_1 z_2 \; u_{y'} e_{y'} D^2_{C/Z}))/ card(V)} \qquad \textbf{(6)}$$

Pour maximiser un tel rapport, on voit qu'il faut privilégier la variable *Y'* qui est bien équilibrée (*i.e.* à modalités de cardinalité proche, pour augmenter $y'_1 y'_2$ qui se trouve au numérateur), dont le produit des probabilités des modalités est faible (afin d'augmenter le contraste des diagrammes de fréquence en diminuant $u_{y'}$ et $e_{y'}$ qui se trouvent bien au dénominateur) et dont la distance des classes décrivant cette variable *Y'* est grande (pour augmenter $D_{C/Y'}$ qui se trouve au numérateur).

Dans le cas où l'on a plus de deux classes, on peut utiliser le critère suivant :

$$\boldsymbol{Sel_4 (Y') = 2\Sigma_{C \in Ck} \Sigma_{Z \in V} y'_1 y'_2 u_z e_z D^2_{C/Y'}/ (1+ z_1 z_2 \; u_{y'} e_{y'} D_{C/Z})^2)/ k(k-1)card(V)}$$

où *k* est le nombre de classes et C_k est l'ensemble des $k(k-1)/2$ couples de classes possibles.

La meilleure variable à sélectionner est alors la variable *Y* qui maximise Sel_4.

Remarquons qu'il serait possible de définir d'autres critères pouvant améliorer encore la qualité du critère de sélection en ajoutant des informations que l'on peut obtenir au niveau *1* des individus précédant le niveau *2* des classes. Pour cela, on peut ajouter par exemple à Sel_4 la valeur du Khi2 ou de la corrélation entre la variable classe *C* et la variable initiale associée à *Y'*. Au niveau *2*, on peut aussi calculer le pouvoir discriminant de *Y'* obtenu par SVM, réseau neuronale, CART etc. Tous ces ajouts doivent être pondérés de façon optimale à l'aide d'algorithmes d'optimisation.

5 Autres formes des critères de sélection de variables symboliques à valeur diagramme en bâtons

5.1 Pouvoir explicatif d'un tableau de données symboliques utilisant la binarité des variables pour prendre en compte le contraste des diagrammes de fréquence à l'aide du critère de Gini

À partir du tableau de données symboliques défini en figure 2 réduit à deux classes,

DC/Y (C1, C2) s'écrit :

$DC/Y\ (C1, C2) = (\Sigma_{j=1,p}\ |Pr(C1 \cap Yj) / Pr(C1) - Pr(C2 \cap Yj) / Pr(C2)\ |)/2).$

On peut définir un premier critère exprimant le pouvoir explicatif de ce tableau, dans le cas réduit à deux classes, par :

$$W_1(Tab_{C/V}) = ((\Sigma_{Y \in V}\ D_{C/Y}\ (C_1, C_2)) / p,$$

où *V* est un ensemble donné de variables symboliques à valeur diagramme en bâtons

et $p = Card(V)$.

En tenant compte du contraste, on pose :

$$D''_{C/Y}\ (C_1, C_2) = D_{C/Y}\ (C_1, C_2)/\ (1+U_Y).$$

On peut alors utiliser le critère suivant : $W_2\ (Tab_{C/V}) = (\Sigma_{Y \in V}\ D''_{C/Y}\ (C_1, C_2))/p.$

On peut de même définir le pouvoir discriminant par :

$$D''_{Y/C}\ (Y_1, Y_2) = D_{Y/C}\ (Y_1, Y_2) / (1+E_Y).$$

D'où de même : $W_2(Tab_{V/C}) = \Sigma_{Y \in V}\ D''_{Y/C}\ (Y_1, Y_2)$, où $Tab_{V/C}$ est le tableau de données symboliques associé à la variable symbolique *C* à valeur « diagramme de fréquence » de modalités Y_i , Z_i, … des variables *Y*, *Z*, etc. de *V*.

Si l'on veut également tenir compte plus ou moins fortement du pouvoir discriminant du tableau $Tab_{Y/C}$, on peut utiliser le critère :

$WG\ (Tab_{C/V}, Tab_{V/C}) =\ a\ \Sigma_{Y \in V}\ D''_{C/Y}\ (C_1, C_2) + b\ \Sigma_{Y \in V}\ D''_{Y/C}(Y_1, Y_2)$ avec $a+b =1$.

Soit : $WG\ (Tab_{C/V}, Tab_{V/C}) =\ a\ W_2(Tab_{C/V}) + b\ W_3(Tab_{V/C})$ avec $a+b =1$.

Il faut utiliser un algorithme d'optimisation pour trouver a et b maximisant WG.

5.2 Les dissimilarités associées à un tableau de données avec plusieurs variables à valeur diagramme en bâtons

Il s'agit de généraliser les dissimilarités utilisées précédemment pour une seule variable au cas de plusieurs variables symboliques. On considère un tableau de données symboliques à *p* variables symboliques à valeur diagrammes définies sur les classes d'une partition *C* à *k* classes. Ces variables sont notées v_j pour *j* variant de *1* à *p* et ces classes sont notées C_i pour *i* variant de *1* à k. Chacune de ces variables v_j a des modalités notées v_{jm} pour *m* variant de *1* à k_j.

Des dissimilarités proportionnelles à l'écart entre $Pr(C_i \cap v_{jm}) / Pr(v_{jm})$ et $Pr(C_{i'} \cap v_{jm}) / Pr(v_{jm})$ peuvent être définies, comme par exemple, pour fixer les idées, la distance L_1 (*i.e.* écart en valeur absolue entre 2 diagrammes en bâtons), ou la distance du Khi2 :

$L_1 :\quad d_{Cii'/vj}\ (C_i, C_{i'}) =\Sigma_{m=1,kj}\ (|Pr(C_i \cap v_{jm}) / Pr(C_i) - Pr(C_{i'} \cap v_{jm}) / Pr(C_{i'})|)/k_j$

où $C_{ii'} = (C_i, C_{i'})$.

Khi2 : $d_{Cii'/vj}(C_i, C_{i'}) = \Sigma_{m=1,kj} |Pr(C_i \cap v_{jm}) / Pr(C_i) - Pr(C_{i'} \cap v_{jm})| / Pr(C_{i'}) | / Pr(v_j)$.

Comme $Pr(C_i) = Pr(C_{i'}) = \Sigma_{j=1,p} k_j$, on peut utiliser :

$$d'_{Cii'/vj}(C_i, C_{i'}) = \Sigma_{m=1,kj} |Pr(C_i \cap v_{jm}) - Pr(C_{i'} \cap v_{jm})| / Pr(v_{jm})$$

à une constante près, ce qui ne modifie donc pas l'ordre des distances.

Remarquons que ces deux distances ont une valeur maximum égale à 1.

On pose : $D_{Cii'/V}(C_i, C_{i'}) = \Sigma_{j=1,p} d_{Cii'/vj}(C_i, C_{i'})/2p$ où p est le nombre de variables Y_j.

$$ID_{C/V} = \Sigma_{i, i'=1,k; i'>i} 2 D_{Cii'/V}(C_i, C_{i'}) / k(k-1).$$

Si l'on considère que P est l'ensemble des couples $C_{ii'} = \{C_i, C_{i'}\}$ avec $i \neq i'$ issus d'une partition des individus $C = \{C1, \ldots, C_k\}$, on peut écrire :

$$ID_{C/V} = \Sigma_{c \in P, v \in V} 2D_{C/V}(c) / k(k-1).$$

Remarquons alors que la valeur maximale de $ID_{C/V}$ est également *1* puisque c'est celle de $D_{Cii'/V}(C_i, C_{i'})$ dont $k(k-1)/2$ termes apparaissent dans cette somme.

5.3 Pouvoir explicatif d'un tableau de données symboliques basé sur la dissimilarité et l'entropie pour prendre en compte le contraste et la "concentration" des diagrammes en bâtons.

L'entropie d'un tableau C/V (voir la figure 2) de données symboliques est définie par :

$$Ent(C/V) = -\Sigma_{i=1,k} \Sigma_{j=1,p} \Sigma_{m=1,kj} Pr(C_i \cap v_{jm}) / Pr(v_{jm}) \, Log\,(Pr(C_i \cap v_j) / Pr(v_{jm})$$

Elle permet de mesurer le degré de contraste des diagrammes de fréquence et aussi de concentration (un tel diagramme est d'autant plus concentré que son nombre de modalités de fréquence faibles ou nulles est grand).

Les cas suivants montrent quelques liens entre entropie et dissimilarité :

i) L'entropie peut être minimum et la dissimilarité minimum (voir Tab_i) ;

ii) L'entropie peut être maximum et la dissimilarité ni minimum ni maximum (voir Tab_{ii}) ;

iii) L'entropie peut être maximum et la dissimilarité maximum (voir Tab_{iii}).

C\ Y	**C\Y₁**	**C\Y₂**
C_1	1	0
C_2	1	0

Tab_i

C\ Y	**C\Y₁**	**C\Y₂**
C_1	2/3	1/3
C_2	1/3	2/3

Tab_{ii}

C\ Y	**C\Y₁**	**C\Y₂**
C_1	1	0
C_2	0	1

Tab_{iii}

Le pouvoir explicatif d'un tableau de données symboliques avec entropie peut s'exprimer sous la forme suivante : $W_{C/V}(Tab_{C/V}) = ID_{C/V} / (1-entr(C/V))$

Ce critère est proportionnel au pouvoir explicatif du tableau *X/Y*. En effet, plus il est grand, plus la somme des distances deux à deux entre les classes C_i est grande et plus l'entropie est faible

(*i.e.* les diagrammes de fréquence sont contrastés). Notons que $W_{C/Y}$ varie entre *0* et *1* puisque la plus grande valeur de $ID_{C/Y}$ est *1* et la plus petite valeur de *-entr(Y/C)* est *0*.

On peut aussi définir de même un critère $C_{Y/C}$ mesurant le pouvoir discriminant du tableau de données symbolique. On peut aussi utiliser un critère mesurant plus ou moins à la fois le pouvoir explicatif et discriminant d'un tableau de données symboliques :

$WE(Tab_{C/V}, Tab_{V/C}) = aW_{C/V}(Tab_{C/V}) + bW_{V/C}(Tab_{V/C})$ avec $a+b = 1$.

5.4 Autres critères de sélection de variables basés sur le critère de Gini et l'entropie

Remarquons que si $p = 1$ (*i.e.* le tableau de données symboliques ne comporte qu'une seule variable) alors *WG* et *WE* constituent deux façons différentes d'exprimer la qualité explicative et discriminante d'une telle variable symbolique.

Plus précisément avec $a+b = 1$.

$WG((Tab_{C/Y}, Tab_{Y/C}) = a\, D_{C/Y}(C_1, C_2)/(1+U_Y) + b\, D_{Y/C}(Y_1, Y_2)/(1+E_Y)$

$WE((Tab_{C/Y}, Tab_{Y/C}) = a\, ID_{C/Y}/(1-entr(C/Y)) + b\, ID_{Y/C}/(1-entr(Y/C))$.

Remarquons néanmoins que *WG* s'applique uniquement dans le cas où la partition *C* comme la variable *Y* sont binaires alors que *WE* peut s'appliquer à plusieurs variables symboliques à valeur diagramme en bâtons ainsi qu'à une partition à plus de deux classes. Notons aussi que Sel_4 à l'avantage de ne pas nécessiter le choix des paramètres *a* et *b* nécessaires pour *WE* et *WG*.

6 Quelques directions de recherche ouvertes

6.1 Construction d'un indicateur logique par arbre à la fois explicatif et discriminant

On note B_Y l'ensemble des variables binaires issues d'une variable Y si elle est formée de plus de 2 modalités. Pour construire un indicateur logique simultanément explicatif et discriminant on peut utiliser le même principe que celui des arbres de décision selon les étapes suivantes :

- choisir la variable $Y_b \Sigma B_Y$ qui maximise Sel_4 ou *WE* puis construire deux segments de façon que

  le premier segment contienne toutes les classes C_i telles que :

  $Pr(Y_1 \cap C_i) > Pr(Y_2 \cap C_i)$

  le second segment contienne alors les autres classes ;

- on peut recommencer le procédé sur chaque segment de classes ainsi obtenu jusqu'à ce que toutes les classes soient séparées. On obtient ainsi un arbre d'explication-décision à la fois explicatif est discriminant des classes initiales. Un nouvel individu peut être associé à un nœud terminal de cet arbre et affecté à la classe la plus fréquente de cet arbre.

On peut ensuite, en se basant sur ce principe, faire du *boosting*, *bagging*, *random forest*, tester la qualité de l'opérateur de classification obtenu, par exemple par *cross validation*, courbe *ROC*, etc. L'intérêt principal par rapport aux méthodes classiques étant d'obtenir un arbre dont les chemins vers les nœuds terminaux font apparaitre les variables qui sont à la fois les plus explicatives et discriminantes des classes les plus fréquentes de ces nœuds terminaux.

6.2 Construction d'un indicateur numérique par combinaison linéaire de variables à valeur diagrammes en bâtons

Pour construire un indicateur numérique, il faut chercher une combinaison linéaire des variables symboliques initiales (supposées binaires ou rendues binaires) qui fournit une nouvelle variable symbolique notée :

$\boldsymbol{Y = \Sigma_{Yj \in V}\, \lambda_j Y_j}$ avec $\boldsymbol{\Sigma_j\, \lambda_j Y_j}$ **qui maximise** le critère Sel_4 ou W, par exemple,

sous la contrainte que *Y* soit une nouvelle variable à valeur diagramme en bâtons. Ceci est toujours possible si le cardinal de *V* est inférieur au nombre de classes. Les λ_j peuvent être obtenus par un algorithme d'optimisation (par exemple, Tabou, Recuit simulé ou Algorithme génétique). La variable *Y* est plus facile à construire quand toutes les variables *Yj* ont le même nombre de modalités et que ces modalités sont ordonnées. Sinon, il faut passer par des stratégies de type « métabins », voir (Diday, 2013). L'intérêt principal par rapport aux méthodes classiques étant d'obtenir un poids d'autant plus fort pour chaque variable qu'elle est explicative et discriminante des classes selon, par exemple, le critère *Sel4* ou *WE* choisi.

6.3 Construction d'une « classification croisée symbolique » basée sur des critères (de sélection de variables croisées) explicatifs et discriminants des classes

Il s'agit de construire une classification croisée des variables qualitatives initiales et des classes d'une partition de façon à obtenir un tableau de données symboliques où chaque classe de la partition recherchée est décrite par un diagramme de fréquence pour chaque classe de variable agrégée par un produit cartésien. On obtient ainsi une « classification croisée symbolique » par opposition à la classification croisée classique qui au lieu de contenir dans chaque case un diagramme en bâtons contient un nombre ou une catégorie. La partition des individus et celle des variables seraient cherchées simultanément par itération successive améliorant à chaque pas un critère *WE*, par exemple. La difficulté viendrait alors du produit cartésien des variables qui deviendrait rapidement complexe avec un nombre de modalités très grand. Par exemple, en croisant dix variables qualitatives à deux modalités on arrive à 210 modalités ! L'idée serait alors de réduire ces croisements en regroupant (par classification automatique), les modalités croisées de profils proches sur les autres variables de façon à obtenir par exemple, un nombre de modalités regroupées en 5 classes. Le tableau symbolique obtenu serait alors défini par une partition des variables et des individus qui contiendrait dans chaque case un diagramme en bâtons à 5 modalités maximisant au mieux un critère à la fois explicatif et discriminant des classes (voir la section 5).

7 Applications industrielles potentielles

Ces résultats s'appliquent par exemple, à la sélection de variables issues de capteurs pour l'aide à la prédiction de classes de niveaux (de confort, de dégradation, de température, de pollution, etc.) compte tenu de conditions environnementales évolutives. En médecine, pour l'étude de trajectoires de patients dans les hôpitaux. En biologie, pour l'étude de l'effet de gènes sur différentes maladies. En marketing, pour l'étude des niveaux de fidélité clientèle. En socio-démographie, pour l'étude et la comparaison de régions ou de pays. En économie, pour trouver de nouveaux indicateurs à la fois explicatifs et discriminants de régions, etc. Dans tous ces cas, on recherche des variables explicatives et discriminantes

Conclusion

Nous avons obtenu plusieurs façons d'exprimer le pouvoir explicatif d'une variable symbolique à valeur diagramme de fréquence. On en a déduit différents critères pour mesurer le pouvoir explicatif d'un tableau de données symboliques. Dans les deux cas, le pouvoir explicatif peut tenir compte à la fois du pouvoir des variables symboliques à discriminer les classes, du pouvoir des classes à discriminer les variables et enfin du contraste des diagrammes de fréquences.

Il reste beaucoup à faire en plus des trois pistes que nous avons suggérées en section 6. D'abord, étudier les propriétés statistiques des différents critères obtenus pouvant déboucher sur des tests de qualité d'une variable ou d'un tableau de données symboliques. Ensuite, il serait intéressant d'étendre ces critères au cas d'autres variables symboliques (à valeur intervalle, histogramme, etc.). Reste aussi à étendre la formulation de certains critères au cas où les diagrammes de fréquence ont plus de deux modalités. Dans ce cas, Gini favorise le contraste mais moins la concentration que l'entropie car il suffit qu'une fréquence s'annule pour que le critère de Gini s'annule alors que l'entropie continue à diminuer. Ces critères peuvent être utilisés pour mettre au point des indicateurs permettant d'expliquer les classes et de les prévoir en probabilité. Ils peuvent aussi être utilisés pour étendre certaines méthodes classiques au cas de données symboliques ou en les utilisant comme en *clustering* (par exemple, pour la décomposition de mélange de lois) ou en classification croisée (*cross-clustering*).

Références

Billard L., Diday E. (2006). *Symbolic Data Analysis: conceptual statistics and data Mining*. Wiley.

Billard L., Diday E. (2003). From the statistics of data to the statistic of knowledge: Symbolic Data Analysis. *JASA - Journal of the American Statistical Association*. Vol. 98, n° 462.

Bock H.H., Diday E. (ed) (2000). *Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg.

Brito P., Bertrand P., Cucumel G., de Carvalho F. (Eds.) (2007). On the Analysis of Symbolic Data. In: *Selected Contributions in Data Analysis and Classification*. Springer, Berlin, pp. 13-22.

Diday E., Noirhomme M. (2008). *Symbolic Data Analysis and the SODAS software*. Wiley.

Diday E. (1987). The symbolic approach in clustering and related methods of data analysis : the basic choices". *First Conference of the International Federation of Classifications Societies*, Technical University of Aachen (RFA).

Diday E. (2013). Principal component analysis for bar charts and metabins tables. *SAM (Statistical Analysis and Data Mining)*, 6(5), pp. 403-430.

Fayyad U., Irani K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *The 12th International Joint Conference on Artificial Intelligence*, pp. 1022–1027.

Fisher W. (1958). On grouping for maximum of homogeneity. *JASA-Journal of the American Statistical Association*.

Noirhomme-Fraiture M., Brito, P. (2012). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining* 4 (2), 157-170.