



HAL
open science

Classification d'un ensemble de variables qualitatives

Hisham Abdallah, Gilbert Saporta

► **To cite this version:**

Hisham Abdallah, Gilbert Saporta. Classification d'un ensemble de variables qualitatives. Revue de Statistique Appliquée, 1998, 46 (4), pp.5-26. hal-02507938

HAL Id: hal-02507938

<https://cnam.hal.science/hal-02507938v1>

Submitted on 16 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

REVUE DE STATISTIQUE APPLIQUÉE

H. ABDALLAH

G. SAPORTA

Classification d'un ensemble de variables qualitatives

Revue de statistique appliquée, tome 46, n° 4 (1998), p. 5-26

http://www.numdam.org/item?id=RSA_1998__46_4_5_0

© Société française de statistique, 1998, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

CLASSIFICATION D'UN ENSEMBLE DE VARIABLES QUALITATIVES

H. Abdallah (1), G. Saporta (2)

(1) Université Pierre et Marie Curie.

Laboratoire de statistique théorique et appliquée, 4 place Jussieu, 75005 Paris.

(2) CNAM, 292 rue Saint-Martin. 75003 Paris.

RÉSUMÉ

Il est fréquent d'avoir à effectuer une typologie d'individus à partir d'un très grand nombre de variables qualitatives. On propose dans cet article des méthodes de classification permettant de remplacer un grand ensemble de variables qualitatives, difficile à gérer, par un ensemble plus réduit.

Mots-clés : Variables qualitatives, analyse relationnelle, classification, sélection, réduction.

ABSTRACT

In many applications using a data base of individual clustering research, we are often led to consider a very important number of qualitative variables. Then the following problem appears : Is it possible to replace a large space of qualitative variables, which is almost impossible to deal with, by a smaller space, containing the same information as the first? The aim of the present article is to solve the problem, classifying qualitative variables with a view to selection and reduction.

Keywords : Qualitative variables, relational analysis, classification, selection, reduction.

1. Introduction

On se donne un ensemble d'individus I décrit par un grand nombre de variables qualitatives $V = \{V_1, \dots, V_m\}$. Notre but est de proposer une méthodologie qui permette de résumer un grand ensemble de variables qualitatives par un petit nombre de variables qualitatives :

– Soit par sélection.

– Soit par constitution d'un petit nombre de variables qualitatives (de façon comparable à l'analyse factorielle qui fournit des variables synthétiques numériques).

Cette méthodologie est différente des méthodes classiques parmi lesquelles :

– L'analyse des correspondances multiples qui consiste à remplacer un grand nombre de variables qualitatives par des coordonnées sur les axes factoriels où ces coordonnées factorielles sont prises comme variables quantitatives pour une classification ultérieure (cf. Saporta [8]).

– La classification d'individus en utilisant le critère de Condorcet qui consiste à partitionner l'ensemble d'individus décrit par l'espace V , en une seule variable qualitative.

Les travaux concernant la classification des variables qualitatives sont peu fréquents dans la littérature et concernent essentiellement les méthodes hiérarchiques (cf. Lerman [4]).

Notre méthodologie se déroule en deux étapes :

– La première étape consiste à classer l'espace des variables qualitatives : Pour cela on introduit de nouveaux critères de similarités entre variables qualitatives ce qui nous permet de résoudre le problème de fixation d'un seuil a priori quand on mesure la liaison entre les variables. Ensuite, on utilise un algorithme de classification de type relationnel qui nous amène à partitionner l'ensemble des variables qualitatives en un certain nombre de classes $J = \{c_1, \dots, c_l, \dots, c_\nu\}$ dont le nombre n est pas fixé a priori.

Deuxième étape :

A ce niveau, deux cas à envisager :

– Dans le cas de sélection de variables qualitatives, on représente chaque classe par la variable qui contribue le plus à la formation de cette classe. Notons $S = \{V_1, \dots, V_l, \dots, V_\nu\}$ l'ensemble des variables sélectionnées.

– Dans le cas de réduction de l'espace des variables qualitatives, on agrège les éléments de chaque classe c_l en un seul élément Y^{*l} (variable consensus). Notons par $J' = \{Y^{*1}, \dots, Y^{*l}, \dots, Y^{*\nu}\}$ l'ensemble des variables agrégées.

Pour valider notre méthodologie dans le dernier cas, on pondère la variable consensus Y^{*l} par le cardinal de la classe qu'elle représente. Notons $R = \{|c_1|Y^{*1}, \dots, |c_\nu|Y^{*\nu}\}$ l'ensemble des variables pondérées. Ensuite on montre qu'on peut borner l'écart entre la partition finale, notée P_R , et celle obtenue sur la totalité des variables, notée P_V .

– Pour mesurer l'accord entre P_R et P_V , il suffit de mesurer l'association entre les deux partitions en utilisant le critère de Rand.

Cette approche est illustrée par l'exemple des « canidés » présenté dans le livre de G. Saporta [8].

2. Études de nouvelles structures d'associations entre variables qualitatives

On construit les tableaux de contingences correspondant au croisement des deux variables qualitatives V_t et $V_{t'}$, ayant respectivement p et q modalités et on mesure alors « l'association » entre ces deux variables, notée $\Omega(V_t, V_{t'})$, en utilisant l'un des critères contingentiels connus : Rand, Chi-Deux, Belson etc..., mais mesurer l'association n'implique pas que les deux variables sont liées ou non, sauf dans le

cas du χ^2 (Test du χ^2). Dans le cas général, comment définir la liaison entre deux variables? Pour cela on définit la dissimilarité $\tilde{\Omega}(V_t, V_{t'})$ entre deux variables V_t et $V_{t'}$ comme le complément de leur similarité $\Omega(V_t, V_{t'})$ à la moyenne de leurs similarités propres $\Omega(V_t, V_t)$ et $\Omega(V_{t'}, V_{t'})$:

$$\tilde{\Omega}(V_t, V_{t'}) = \frac{\Omega(V_t, V_t) + \Omega(V_{t'}, V_{t'})}{2} - \Omega(V_t, V_{t'}).$$

Ensuite, nous proposons, et afin d'éviter le calcul de seuils probabilistes, un critère $H_{tt'}$, basé sur la comparaison de $\Omega(V_t, V_{t'})$ par rapport à $\tilde{\Omega}(V_t, V_{t'})$, défini par :

$$H_{tt'} = \Omega(V_t, V_{t'}) - \tilde{\Omega}(V_t, V_{t'}) = 2\Omega(V_t, V_{t'}) - \frac{\Omega(V_t, V_t) + \Omega(V_{t'}, V_{t'})}{2}.$$

On dit que les deux variables sont liées si $H_{tt'} \geq 0$ c'est-à-dire :

$$\Omega(V_t, V_{t'}) \geq \tilde{\Omega}(V_t, V_{t'}).$$

Nous ne prétendons pas donner un inventaire global des critères usuels d'associations, mais nous exposerons ceux qui nous paraissent importants pour notre propos, en particulier, nous présentons des critères peu connus mais qui ont une grande utilité comme mesure d'association.

Tous les critères donnés ci-dessous seront définis contingemment, mais on aurait pu aussi les exprimer de façon relationnelle (cf Marcotorchino [5], Marcotorchino-Michaud [6]).

2.1 Critère de Rand

Ce critère a été introduit en 1971 par Rand sous une forme très légèrement différente de celle que nous donnons ici, mais totalement équivalente. Sur un tableau de contingence T, $R(V_t, V_{t'})$ s'écrit dans sa forme généralisée donnée dans [5].

$$R(V_t, V_{t'}) = \frac{2 \sum_{u=1}^p \sum_{\nu=1}^q n_{u\nu}^2 - \sum_{u=1}^p n_{u.}^2 - \sum_{\nu=1}^q n_{.\nu}^2 + n^2}{n^2}$$

où :

$n_{u\nu}$: nombre d'individus ayant la modalité u de V_t et la modalité ν de $V_{t'}$.

$n_{u.}$: nombre d'individus ayant la modalité u de V_t .

$n_{.\nu}$: nombre d'individus ayant la modalité ν de $V_{t'}$.

p : nombre de modalités de V_t .

q : nombre de modalités de $V_{t'}$.

Définissons $\tilde{R}(V_t, V_{t'})$ par :

$$\tilde{R}(V_t, V_{t'}) = \frac{R(V_t, V_t) + R(V_{t'}, V_{t'})}{2} - R(V_t, V_{t'})$$

Comme il est évident que $R(V_t, V_t) = 1$, on trouve :

$$\tilde{R}(V_t, V_{t'}) = \frac{\sum_{u=1}^p n_u^2 + \sum_{\nu=1}^q n_\nu^2 - 2 \sum_{u=1}^p \sum_{\nu=1}^q n_{u\nu}^2}{n^2}$$

d'où :

$$\begin{aligned} H_{tt'}^1 &= R(V_t, V_{t'}) - \tilde{R}(V_t, V_{t'}) \\ &= 2 \frac{\sum_{u\nu} n_{u\nu}^2 - \sum_u n_u^2 - \sum_\nu n_\nu^2 + n^2}{n^2} - 1 \\ &= 2R(V_t, V_{t'}) - 1 \end{aligned}$$

On retrouve ici que $H_{tt'}^1 \geq 0$, dès que $R(V_t, V_{t'}) \geq \frac{1}{2}$. (Sachant par ailleurs que $0 \leq R(V_t, V_{t'}) \leq 1$ pour tous t et t').

2.2 Critère de l'écart à l'indétermination

Il s'écrit contingentiellement (cf. Marcotorchino [5]) :

$$I(V_t, V_{t'}) = \sum_{u,\nu} \left[n_{u\nu} - \left(\frac{n_u}{q} + \frac{n_\nu}{p} - \frac{n}{pq} \right) \right]^2$$

Ce qui s'écrit encore en développant et en simplifiant :

$$I(V_t, V_{t'}) = \sum_{u\nu} n_{u\nu}^2 - \left(\sum_u \frac{n_u^2}{q} + \sum_\nu \frac{n_\nu^2}{p} \right) + \frac{n^2}{pq}$$

De même, définissons $\tilde{I}(V_t, V_{t'})$ par :

$$\tilde{I}(V_t, V_{t'}) = \frac{I(V_t, V_t) + I(V_{t'}, V_{t'})}{2} - I(V_t, V_{t'})$$

avec

$$I(V_t, V_t) = \sum_u n_u^2 \left(1 - \frac{2}{p} \right) + \frac{n^2}{p^2}, \quad I(V_{t'}, V_{t'}) = \sum_\nu n_\nu^2 \left(1 - \frac{2}{q} \right) + \frac{n^2}{q^2}$$

d'où :

$$\begin{aligned}
H_{tt'}^2 &= I(V_t, V_{t'}) - \tilde{I}(V_t, V_{t'}) \\
&= 2I(V_t, V_{t'}) - \frac{I(V_t, V_t) + I(V_{t'}, V_{t'})}{2} \\
&= 2 \sum_{u, \nu} n_{u\nu}^2 - \sum_u n_u^2 \left(\frac{2}{q} + \frac{1}{2} - \frac{1}{p} \right) - \sum_\nu n_\nu^2 \left(\frac{2}{p} + \frac{1}{2} - \frac{1}{q} \right) \\
&\quad + n^2 \left(\frac{2}{pq} - \frac{1}{2p^2} - \frac{1}{2q^2} \right) \\
&= n^2 R(V_t, V_{t'}) - n^2 + \sum_u n_u^2 \left(\frac{1}{2} - \frac{2}{q} + \frac{1}{p} \right) \\
&\quad + \sum_\nu n_\nu^2 \left(\frac{1}{q} - \frac{2}{p} + \frac{1}{2} \right) + n^2 \left(\frac{2}{pq} - \frac{1}{2p^2} - \frac{1}{2q^2} \right)
\end{aligned}$$

Dire que $H_{tt'}^2 \geq 0$ implique que

$$\begin{aligned}
R(V_t, V_{t'}) \geq 1 - \frac{1}{n^2} \left[\sum_u n_u^2 \left(\frac{1}{2} + \left(\frac{1}{p} - \frac{2}{q} \right) \right) + \sum_\nu n_\nu^2 \left(\frac{1}{2} + \left(\frac{1}{q} - \frac{2}{p} \right) \right) \right] \\
+ \frac{1}{2p^2} + \frac{1}{2q^2} - \frac{2}{pq} = A.
\end{aligned}$$

A se met sous la forme suivante : $1 - \frac{1}{n^2} \left(\alpha \sum_u n_u^2 + \beta \sum_\nu n_\nu^2 \right) + \gamma$ avec

$$\alpha = \frac{1}{2} + \frac{1}{p} - \frac{2}{q}; \beta = \frac{1}{2} + \frac{1}{q} - \frac{2}{p}; \gamma = \frac{1}{2p^2} + \frac{1}{2q^2} - \frac{2}{pq}.$$

Nous savons que

$$\frac{n^2}{p} \leq \sum_u n_u^2 \leq (n - (p - 1))^2 + p - 1$$

et

$$\frac{n^2}{q} \leq \sum_\nu n_\nu^2 \leq (n - (q - 1))^2 + q - 1$$

Nous allons déterminer les bornes de A à partir de celles de $\sum_u n_u^2$ et celles de $\sum_\nu n_\nu^2$, et le fait que α et β n'ont pas un signe constant quand p et q varient, on va étudier les 3 cas suivants :

$\alpha \geq 0$ et $\beta \geq 0$;

$\alpha \leq 0$ et $\beta \geq 0$; $\alpha \geq 0$ et $\beta \leq 0$ (le cas $\alpha < 0$, $\beta < 0$, ne peut pas se produire, p et q étant supérieurs ou égaux à 2).

1^{er} cas : pour $\alpha \geq 0$ et $\beta \geq 0$, on a :

$$\begin{aligned} A &\leq 1 - \frac{1}{n^2} \left[\frac{n^2}{p} \left(\frac{1}{2} + \frac{1}{p} - \frac{2}{q} \right) + \frac{n^2}{q} \left(\frac{1}{2} + \frac{1}{q} - \frac{2}{p} \right) \right] + \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 - \frac{1}{pq} \\ &\leq 1 - \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 - \frac{1}{2} \left(\frac{1}{p} + \frac{1}{q} \right) + \frac{1}{pq}. \end{aligned}$$

2^{ème} cas : pour $\alpha \geq 0$ et $\beta \leq 0$, on a :

$$A \leq 1 + \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 - \frac{1}{pq} - \frac{\alpha}{p} - \beta + O\left(\frac{1}{n}\right) f(\beta, q)$$

si n est grand, on a :

$$A \leq \frac{1}{2} + \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 + \frac{1}{pq} - \frac{1}{p^2} - \frac{1}{q} + \frac{3}{2p}$$

3^{ème} cas : pour $\alpha \leq 0$ et $\beta \geq 0$, on a :

$$A \leq 1 + \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 - \frac{1}{pq} - \frac{\beta}{q} - \alpha + O\left(\frac{1}{n}\right) f(\alpha, p)$$

si n est grand, on a :

$$A \leq \frac{1}{2} + \frac{1}{2} \left(\frac{1}{p} - \frac{1}{q} \right)^2 + \frac{1}{pq} - \frac{1}{q^2} - \frac{1}{p} + \frac{3}{2q}$$

Si donc $R(V_t, V_{t'})$ est supérieur ou égale à la borne trouvée pour A (et qui dépend comme on vient de le voir du signe de α et de β) alors $R(V_t, V_{t'}) \geq A$, et donc $H_{tt'}^2 \geq 0$.

Cas particuliers :

Si $p = q$.

La forme de $H_{tt'}^2$ devient dans ce cas :

$$\begin{aligned} H_{tt'}^2 &= 2 \sum_{u, \nu} n_{u\nu}^2 - \sum_u n_u^2 \left(\frac{1}{p} + \frac{1}{2} \right) - \sum_\nu n_\nu^2 \left(\frac{1}{p} + \frac{1}{2} \right) + \frac{n^2}{p^2} \\ &= n^2 R(V_t, V_{t'}) \left(\frac{1}{2} + \frac{1}{p} \right) + 2 \sum_{u, \nu} n_{u\nu}^2 \left(\frac{1}{2} - \frac{1}{p} \right) - n^2 \left(\frac{1}{2} + \frac{1}{p} - \frac{1}{p^2} \right) \end{aligned}$$

pour plus de détail voir [1]. Dire que $H_{tt'}^2 \geq 0$, revient à dire que :

$$R(V_t, V_{t'}) \geq \frac{2}{n^2} \left(\frac{2-p}{2+p} \right) \sum_{u,\nu} n_{u\nu}^2 + \frac{p^2 + 2p - 2}{p(p+2)}$$

Si $p = 2$.

Dire que $H_{tt'}^2 \geq 0$ dans ce cas implique : $R(V_t, V_{t'}) \geq \frac{3}{4}$.

2.3 Critère de l'écart carré à l'indépendance

Il s'écrit dans sa formulation contingentielle :

$$E(V_t, V_{t'}) = \sum_{u=1}^p \sum_{\nu=1}^q \left(n_{u\nu}^2 - \frac{n_u^2 \cdot n_\nu^2}{n^2} \right)$$

Définissons $\tilde{E}(V_t, V_{t'})$ par :

$$\tilde{E}(V_t, V_{t'}) = \frac{E(V_t, V_t) + E(V_{t'}, V_{t'})}{2} - E(V_t, V_{t'})$$

avec :

$$E(V_t, V_t) = \sum_u n_u^2 - \frac{\left(\sum_u n_u^2 \right)^2}{n^2}$$

$$E(V_{t'}, V_{t'}) = \sum_\nu n_\nu^2 - \frac{\left(\sum_\nu n_\nu^2 \right)^2}{n^2}$$

d'où :

$$\begin{aligned} H_{tt'}^3 &= E(V_t, V_{t'}) - \tilde{E}(V_t, V_{t'}) \\ &= 2E(V_t, V_{t'}) - \frac{E(V_t, V_t) + E(V_{t'}, V_{t'})}{2} \\ &= 2 \sum_{u\nu} \left(n_{u\nu}^2 - \frac{n_u^2 \cdot n_\nu^2}{n^2} \right) - \frac{1}{2} \left(\sum_u n_u^2 + \sum_\nu n_\nu^2 \right) \\ &\quad + \frac{1}{2n^2} \left(\left(\sum_u n_u^2 \right)^2 + \left(\sum_\nu n_\nu^2 \right)^2 \right) \end{aligned}$$

Ce qui s'écrit encore :

$$\begin{aligned}
 H_{tt'}^3 &= 2 \sum_{u,\nu} n_{u\nu}^2 - \frac{1}{2} \left[\left(\sum_u n_u^2 + \sum_\nu n_\nu^2 \right) - \frac{\left(\sum_u n_u^2 - \sum_\nu n_\nu^2 \right)^2}{n^2} \right] \\
 &\quad - \frac{\sum_u n_u^2 \sum_\nu n_\nu^2}{n^2} \\
 &= n^2 R(V_t, V_{t'}) - n^2 + \frac{\left(\sum_u n_u^2 - \sum_\nu n_\nu^2 \right)^2}{2n^2} + \frac{1}{2} \sum_u n_u^2 \\
 &\quad + \frac{1}{2} \sum_\nu n_\nu^2 - \frac{\sum_u n_u^2 \sum_\nu n_\nu^2}{n^2}
 \end{aligned}$$

Soit $H_{tt'}^3 \geq 0$ ceci implique que :

$$R(V_t, V_{t'}) \geq 1 - \left[\frac{\left(\sum_u n_u^2 - \sum_\nu n_\nu^2 \right)^2}{2n^4} + \frac{\sum_u n_u^2 + \sum_\nu n_\nu^2}{2n^2} \right] + \frac{\sum_u n_u^2 \sum_\nu n_\nu^2}{n^4} = B$$

Et si l'on prend les bornes de $\sum_u n_u^2$ et celles de $\sum_\nu n_\nu^2$ déjà citées ci-dessus, on trouve :

$$\begin{aligned}
 B &\leq 1 + \frac{2}{n^4} ((n-p+1)^2 + p-1)((n-q+1)^2 + q-1) \\
 &\quad - \frac{1}{2} \left(\frac{1}{p^2} + \frac{1}{q^2} + \frac{1}{p} + \frac{1}{q} \right) = B_0
 \end{aligned}$$

Si donc $R(V_t, V_{t'})$ est supérieur ou égale à B_0 , $R(V_t, V_{t'})$ est supérieur ou égal à B et donc $H_{tt'}^3 \geq 0$.

2.4 Critère du Chi-deux

Il est défini par la formule suivante :

$$\chi^2(V_t, V_{t'}) = n \left(\sum_{u,\nu} \frac{n_{u\nu}^2}{n_u n_\nu} - 1 \right) = n\phi^2(V_t, V_{t'})$$

Définissons de même $\tilde{\chi}^2(V_t, V_{t'})$ par :

$$\tilde{\chi}^2(V_t, V_{t'}) = \frac{\chi^2(V_t, V_t) + \chi^2(V_{t'}, V_{t'})}{2} - \chi^2(V_t, V_{t'})$$

avec

$$\begin{aligned}\chi^2(V_t, V_t) &= n(p-1) \\ \chi^2(V_{t'}, V_{t'}) &= n(q-1)\end{aligned}$$

donc

$$\begin{aligned}\tilde{\chi}^2(V_t, V_{t'}) &= \frac{n(p-1) + n(q-1)}{2} - n \left(\sum_{uv} \frac{n_{uv}^2}{n_u \cdot n_v} - 1 \right) \\ &= n \left(\frac{p+q}{2} - \sum_{uv} \frac{n_{uv}^2}{n_u \cdot n_v} \right)\end{aligned}$$

d'où :

$$\begin{aligned}H_{tt'}^4 &= \chi^2(V_t, V_{t'}) - \tilde{\chi}^2(V_t, V_{t'}) \\ &= n \left(2 \sum_{u,v} \frac{n_{uv}^2}{n_u \cdot n_v} - \frac{p+q}{2} - 1 \right) \\ &= n \left(2\phi^2(V_t, V_{t'}) - \frac{p+q-2}{2} \right)\end{aligned}$$

Dire que $H_{tt'}^4 \geq 0$ implique que

$$\phi^2(V_t, V_{t'}) \geq \frac{p+q-2}{4}$$

Remarque 1.

Il existe une version «indice» d'association normée entre 0 et 1, notée χ_{norm}^2 , donnée par Cramer (1946) sous la forme suivante :

$$\chi_{\text{norm}}^2(V_t, V_{t'}) = \frac{\chi^2}{n \text{ Min} [(p-1), (q-1)]} = \frac{\sum_{u,v} \frac{n_{uv}^2}{n_u \cdot n_v} - 1}{\text{Min} [(p-1), (q-1)]}$$

avec :

$$\begin{aligned}\chi_{\text{norm}}^2(V_t, V_t) &= 1 \\ \chi_{\text{norm}}^2(V_{t'}, V_{t'}) &= 1\end{aligned}$$

donc :

$$\tilde{\chi}_{\text{norm}}^2(V_t, V_{t'}) = 1 - \frac{\sum_{u,v} \frac{n_{uv}^2}{n_u \cdot n_v} - 1}{\text{Min}[(p-1), (q-1)]}$$

On définit alors :

$$H_{tt'}'^4 = 2\chi_{\text{norm}}^2(V_t, V_{t'}) - 1$$

et on retrouve ici que $H_{tt'}'^4 \geq 0$, dès que $\chi_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$.

Une autre normalisation utilisée est celle du coefficient de Tchuprow, qui consiste à diviser le ϕ^2 par la moyenne géométrique de $(p-1)$ et $(q-1)$.

d'où :

$$\phi_{\text{norm}}^2 = \frac{\sum_{u,v} \frac{n_{uv}^2}{n_u \cdot n_v} - 1}{\sqrt{(p-1)(q-1)}}$$

avec :

$$\phi_{\text{norm}}^2(V_t, V_t) = 1 = \phi_{\text{norm}}^2(V_{t'}, V_{t'})$$

D'où

$$\tilde{\phi}_{\text{norm}}^2(V_t, V_{t'}) = 1 - \frac{\sum_{u,v} \frac{n_{uv}^2}{n_u \cdot n_v} - 1}{\sqrt{(p-1)(q-1)}}$$

On définit alors :

$$H_{tt'}''^4 = 2\tilde{\phi}_{\text{norm}}^2(V_t, V_{t'}) - 1$$

Dire que $H_{tt'}''^4 \geq 0$ implique que $\tilde{\phi}_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$.

2.5 Critère dérivé de Jordan

Il s'écrit contingentiellement :

$$J(V_t, V_{t'}) = \sum_{u,v} \frac{n_{uv}}{n} \left(n_{uv} - \frac{n_u \cdot n_v}{n} \right)$$

De même définissons $\tilde{J}(V_t, V_{t'})$ par :

$$\tilde{J}(V_t, V_{t'}) = \frac{J(V_t, V_t) + J(V_{t'}, V_{t'})}{2} - J(V_t, V_{t'})$$

avec :

$$J(V_t, V_t) = \frac{1}{n} \sum_u \left(n_{u.}^3 - \frac{n_{u.}^3}{n} \right)$$

$$J(V_{t'}, V_{t'}) = \frac{1}{n} \sum_\nu \left(n_{.\nu}^3 - \frac{n_{.\nu}^3}{n} \right)$$

Compte tenu de ce que $\sum_u n_{u.}^3 = \sum_{u\nu} n_{u\nu} n_{u.}^2$, et $\sum_\nu n_{.\nu}^3 = \sum_{u\nu} n_{u\nu} n_{.\nu}^2$, on a :

$$\begin{aligned} H_{tt'}^5 &= J(V_t, V_t) - \tilde{J}(V_t, V_{t'}) \\ &= \frac{2}{n} \sum_{u,\nu} n_{u\nu}^2 - \frac{1}{2n} \left[\sum_u n_{u.}^2 + \sum_\nu n_{.\nu}^2 \right] \\ &\quad + \frac{1}{2n^2} \sum_{u,\nu} n_{u\nu} n_{u.}^2 + \frac{1}{2n^2} \sum_{u,\nu} n_{u,\nu} n_{.\nu}^2 - \frac{2}{n^2} \sum_{u,\nu} n_{u\nu} (n_{u.} n_{.\nu}) \\ &= \frac{1}{n} [n^2 R(V_t, V_{t'})] - n + \frac{1}{2n} \left[\sum_u n_{u.}^2 + \sum_\nu n_{.\nu}^2 \right] \\ &\quad + \frac{1}{2n^2} \left[\sum_{u,\nu} n_{u\nu} (n_{u.} - n_{.\nu})^2 \right] - \frac{1}{n^2} \sum_{u,\nu} n_{u\nu} n_{u.} n_{.\nu}. \end{aligned}$$

Dire que $H_{tt'}^5 \geq 0$ implique que :

$$R(V_t, V_{t'}) \geq 1 - \alpha_1$$

où

$$\alpha_1 = \frac{1}{2n^2} \left[\sum_u n_{u.}^2 + \sum_\nu n_{.\nu}^2 \right] + \frac{1}{2n^3} \left[\sum_{u,\nu} n_{u\nu} (n_{u.} - n_{.\nu})^2 \right] - \frac{1}{n^3} \sum_{u,\nu} n_{u\nu} n_{u.} n_{.\nu}$$

et si l'on prend les bornes de $\sum_u n_{u.}^2$ et celles de $\sum_\nu n_{.\nu}^2$, on a :

$$1 - \alpha_1 \leq 1 - \frac{1}{2} \left(\frac{1}{p} + \frac{1}{q} \right) - \frac{1}{2n^3} \left[\sum_{u,\nu} n_{u\nu} (n_{u.} - n_{.\nu})^2 \right] + \frac{1}{n^3} \sum_{u,\nu} n_{u\nu} n_{u.} n_{.\nu}$$

Si donc $R(V_t, V_{t'})$ est supérieur ou égale à la quantité figurant au second membre de l'expression précédente, on a $R(V_t, V_{t'}) \geq 1 - \alpha_1$ et donc $H_{tt'}^5 \geq 0$.

2.6 Critère de Belson

Il s'écrit contingentiellement :

$$B(V_t, V_{t'}) = \sum_{u,\nu} \left(n_{u\nu} - \frac{n_u \cdot n_\nu}{n} \right)^2$$

Définissons $\tilde{B}(V_t, V_{t'})$ par :

$$\tilde{B}(V_t, V_{t'}) = \frac{B(V_t, V_t) + B(V_{t'}, V_{t'})}{2} - B(V_t, V_{t'})$$

avec :

$$B(V_t, V_t) = \sum_u n_u^2 - \frac{2}{n} \sum_u n_u^3 + \frac{\left(\sum_u n_u^2 \right)^2}{n^2}$$

$$B(V_{t'}, V_{t'}) = \sum_\nu n_\nu^2 - \frac{2}{n} \sum_\nu n_\nu^3 + \frac{\left(\sum_\nu n_\nu^2 \right)^2}{n^2}$$

Tenant compte des relations $\sum_u n_u^3 = \sum_{u\nu} n_{u\nu} n_u^2$, et $\sum_\nu n_\nu^3 = \sum_{u\nu} n_{u\nu} n_\nu^2$, déjà utilisées ci-dessus, on a :

$$\begin{aligned} H_{tt'}^6 &= B(V_t, V_{t'}) - \tilde{B}(V_t, V_{t'}) \\ &= 2 \sum_{u,\nu} n_{u\nu}^2 - \frac{1}{2} \left[\sum_u n_u^2 + \sum_\nu n_\nu^2 \right] - \frac{4}{n} \sum_{u,\nu} n_{u\nu} n_u \cdot n_\nu \\ &\quad - \frac{\left(\sum_u n_u^2 \right)^2}{2n^2} - \frac{\left(\sum_\nu n_\nu^2 \right)^2}{2n^2} \\ &\quad + \frac{2}{n^2} \sum_u n_u^2 \cdot \sum_\nu n_\nu^2 + \frac{1}{n} \left[\sum_{u,\nu} n_{u\nu} n_u^2 + \sum_{u,\nu} n_{u\nu} n_\nu^2 \right] \\ &= n^2 R(V_t, V_{t'}) - n^2 + \frac{1}{2} \left[\sum_u n_u^2 + \sum_\nu n_\nu^2 \right] - \frac{\left(\sum_u n_u^2 \right)^2}{2n^2} - \frac{\left(\sum_\nu n_\nu^2 \right)^2}{2n^2} \\ &\quad + \frac{2}{n^2} \sum_u n_u^2 \cdot \sum_\nu n_\nu^2 + \frac{1}{n} \left[\sum_{u,\nu} n_{u\nu} (n_u - n_\nu)^2 \right] - \frac{2}{n} \sum_{u\nu} n_{u\nu} n_u \cdot n_\nu \end{aligned}$$

Dire que $H_{tt'}^6 \geq 0$ implique que :

$$R(V_t, V_{t'}) \geq 1 - \alpha_2$$

où

$$\begin{aligned} \alpha_2 = & \frac{1}{2n^2} \left[\sum_u n_u^2 + \sum_\nu n_\nu^2 \right] - \frac{\left(\sum_u n_u^2 \right)^2}{2n^4} - \frac{\left(\sum_\nu n_\nu^2 \right)^2}{2n^4} \\ & + \frac{2}{n^4} \sum_u n_u^2 \cdot \sum_\nu n_\nu^2 + \frac{1}{n^3} \left[\sum_{u,\nu} n_{u\nu} (n_u - n_\nu)^2 \right] - \frac{2}{n^3} \sum_{u\nu} n_{u\nu} n_u \cdot n_\nu. \end{aligned}$$

De même, si l'on prend les bornes $\sum_u n_u^2$ et celles de $\sum_\nu n_\nu^2$ définies ci-dessus, on a :

$$\begin{aligned} 1 - \alpha_2 \leq & 1 - \frac{1}{2} \left(\frac{1}{p} + \frac{1}{q} \right) - \frac{2}{pq} + \frac{((n-p+1)^2 + p-1)^2}{2n^4} \\ & + \frac{((n-q+1)^2 + q-1)^2}{2n^4} - \frac{1}{n^3} \left[\sum_{u,\nu} n_{u\nu} (n_u - n_\nu)^2 \right] + \frac{2}{n^3} \sum_{u\nu} n_{u\nu} n_u \cdot n_\nu. \end{aligned}$$

Si donc $R(V_t, V_{t'})$ est supérieur ou égale à la quantité figurant au second membre de l'expression précédente, on a $R(V_t, V_{t'}) \geq 1 - \alpha_2$ et donc $H_{tt'}^6 \geq 0$.

2.7 Synthèse

D'après ce qui précède, nous avons montré que $H_{tt'}^1, H_{tt'}^2, H_{tt'}^3, H_{tt'}^5$ et $H_{tt'}^6$ s'expriment en fonction de $R(V_t, V_{t'})$, et $H_{tt'}^4, H_{tt'}'^4$ et $H_{tt'}''^4$ s'expriment en fonction de $\phi^2(V_t, V_{t'})$.

On en déduit que le seuil de classification des variables, selon $H_{tt'}^2$ et $H_{tt'}^3$, dépend des valeurs attribuées à p et q , et celui, selon $H_{tt'}^1, H_{tt'}'^4, H_{tt'}''^4$ est fixe et égal à 0.5.

Le tableau ci-après présente, sous forme synthétique, les cas correspondant à $\Omega(V_t, V_{t'})$.

On retrouve que les critères $H_{tt'}^1, H_{tt'}^4, H_{tt'}'^4$ et $H_{tt'}''^4$ sont positifs dès qu'on a respectivement $R(V_t, V_{t'}) \geq \frac{1}{2}$, $\phi^2(V_t, V_{t'}) \geq \frac{p+q-2}{4}$, $\chi_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$ et $\phi_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$.

$H_{tt'} = \Omega(V_t, V_{t'}) - \tilde{\Omega}(V_t, V_{t'})$	Implication au niveau de la borne de référence maximale.
<p>Critère de Rand :</p> $R(V_t, V_{t'})$. $H_{tt'}^1 = R(V_t, V_{t'}) - \tilde{R}(V_t, V_{t'})$	$R(V_t, V_{t'}) \geq \frac{1}{2}$
<p>Critère de l'écart à l'indétermination :</p> $I(V_t, V_{t'})$. $H_{tt'}^2 = I(V_t, V_{t'}) - \tilde{I}(V_t, V_{t'})$	dépend des valeurs de p et q Cas particulier : $p = q = 2$; $R(V_t, V_{t'}) \geq \frac{3}{4}$
<p>Critère de l'écart à l'indépendance :</p> $E(V_t, V_{t'})$. $H_{tt'}^3 = E(V_t, V_{t'}) - \tilde{E}(V_t, V_{t'})$	dépend des valeurs de p et q
$\chi^2(V_t, V_{t'})$ $H_{tt'}^4 = \chi^2(V_t, V_{t'}) - \tilde{\chi}^2(V_t, V_{t'})$	$\phi^2(V_t, V_{t'}) \geq \frac{p+q-2}{4}$
$\chi_{\text{norm}}^2(V_t, V_{t'})$ $H_{tt'}'^4 = \chi_{\text{norm}}^2(V_t, V_{t'}) - \tilde{\chi}_{\text{norm}}^2(V_t, V_{t'})$ $\phi_{\text{norm}}^2(V_t, V_{t'})$ $H_{tt'}''4 = \phi_{\text{norm}}^2(V_t, V_{t'}) - \tilde{\phi}_{\text{norm}}^2(V_t, V_{t'})$	$\chi_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$ $\phi_{\text{norm}}^2(V_t, V_{t'}) \geq \frac{1}{2}$
<p>critère dérivé de Jordan :</p> $J(V_t, V_{t'})$ $H_{tt'}^5 = J(V_t, V_{t'}) - \tilde{J}(V_t, V_{t'})$	dépend de p et de q
<p>critère de Belson :</p> $B(V_t, V_{t'})$ $H_{tt'}^6 = B(V_t, V_{t'}) - \tilde{B}(V_t, V_{t'})$	dépend de p et de q

3. Partition de l'ensemble des variables qualitatives

D'après ce qui précède, le problème de classification de l'ensemble des m variables dans l'optique relationnelle consiste à chercher une partition de l'ensemble V en classes d'équivalences, c'est-à-dire à chercher une relation d'équivalence que nous appellerons Y , qui maximisera notre critère.

Sachant que la relation inconnue Y est représentée par une matrice booléenne de terme général $y_{tt'}$, vérifiant les contraintes linéaires de symétrie et de transitivité, maximiser notre critère, revient à résoudre le programme en variables (0-1) suivant :

$$\max_Y \sum_{t=1}^m \sum_{t'=1}^m H_{tt'} y_{tt'}$$

avec : $y_{tt'} - y_{t't} = 0 \forall t, t'$ (symétrie).

$y_{tt'} + y_{t't''} - y_{tt''} \leq 1 \forall (t, t', t'')$ (transitivité).

$y_{tt'} \in \{0, 1\}$ (binarité).

La solution Y de ce problème est obtenue par l'utilisation d'une méthode heuristique, (celle du C.E.M.A.P., IBM).

La solution ainsi trouvée est une partition de l'ensemble des variables en ν classes non empiétantes dont on n'a fixé a priori ni la taille, ni le nombre, contrairement à ce qui se produit communément en classification.

Notons $J = \{c_1, c_2, \dots, c_\nu\}$ la partition de l'ensemble des variables en ν classes, ou $\bigcup c_l = V$ (c_l peut être composée d'un élément unique ou de plusieurs éléments) et $c_l \cap c_j = \{\emptyset\}$; $l = 1, \dots, \nu$ et $j = 1, \dots, \nu, l \neq j$.

La dimension de J représentera le nombre optimal des classes obtenues ($\dim J < \dim V$). Chaque élément de J est une nouvelle variable synthétique représentant les variables initiales qui ont été regroupées ensemble dans cette classe. Cette première étape a pour objet d'exhiber l'information contenue dans l'ensemble des variables sous forme de quelques variables synthétiques.

Cette étape de classification relationnelle des variables nous a fourni des variables qui structurent et résument au mieux les variables initiales. Dans cette classification, deux variables ont d'autant plus de chance d'être regroupées que le profil des individus qu'ils décrivent est semblable.

A ce niveau, deux cas sont à envisager :

3.1. 1^{er} cas : Sélection de variables qualitatives

On représente chaque classe par la variable qui contribue le plus à la formation de cette classe, la contribution est donnée par la formule suivante :

$$\forall V_t \in c_l, C(V_t, c_l) = 2 \sum_{V_{t'} \in c_l} H_{tt'}$$

Cela nous conduit à sélectionner un nombre de variables égal au nombre de classes obtenues.

Notons $S = \{V_1, \dots, V_l, \dots, V_\nu\}$ l'ensemble des variables sélectionnées.

3.2. 2^{ème} cas : Réduction de l'espace des variables qualitatives

Dans ce cas, on résume l'ensemble des variables appartenant à chaque classe c_l de J par une nouvelle variable qualitative Y^{*l} dont les modalités sont définies par les classes de la partition sur les individus obtenue en utilisant la méthode de Condorcet. Si la classe c_l a un unique élément on a $Y^{*l} = V_l$.

Soit l l'ensemble des variables initiales d'une classe c_l de J . Partant alors d'un tableau $T(n, l)$, où chaque colonne du tableau représente l'ensemble des modalités données par une variable aux n individus. Citons par exemple, $c_l = \{V_1, V_5, V_{12}\}$, T représente la description des individus relativement à V_1, V_5, V_{12} .

On se propose de trouver le vecteur inconnu à n dimensions Y^{*l} donné par une variable agrégeant l'ensemble de ces variables.

La matrice de base est la matrice C^l des comparaisons par paires entre individus pour chaque classe appelée matrice de Condorcet. Son terme général $c_{ii'}^l$, est défini par :

$c_{ii'}^l$: nombre de variables de la classe c_l qui réunissent les individus i et i' .

On cherche donc à regrouper, au sein de classes homogènes, les individus qui s'apparentent tant par les ressemblances qu'ils entretiennent que par leurs oppositions aux autres individus.

Le problème se pose sous la forme de ν modèles linéaires sous contraintes suivants :

$$\max_Y \sum_i \sum_{i \neq i'} c_{ii'}^l Y_{ii'}^{*l} \quad (1)$$

où $l = 1, \dots, \nu$

$$\begin{aligned} Y_{ii'}^{*l} - Y_{i'i}^{*l} &= 0 \quad i \neq i' \\ Y_{ii'}^{*l} + Y_{i'i''}^{*l} - Y_{ii''}^{*l} &\leq 1, \quad \text{pour tout } i \neq i', i' \neq i'', i \neq i'' \\ Y_{ii'}^{*l} &\in \{0, 1\} \end{aligned}$$

où $c_{ii'}^l$ représente le nombre de variables de la classe c_l , pour lesquels les individus i et i' ont la même modalité, moins le nombre de variables pour lesquels i et i' n'ont pas la même modalité pour la même classe. De plus (1) peut s'écrire de la façon équivalente suivante :

$$\text{Max} \sum_{i, i'} \sum_{l=1}^{\nu} \sum_{t \in c_l} (c_{ii'}^t Y_{ii'}^{*l} + \tilde{c}_{ii'}^t \tilde{Y}_{ii'}^{*l}).$$

avec $c_{ii'}^t = 1$ si i et i' ont adopté la même modalité de V_t , et $c_{ii'}^t = 0$ sinon et $\tilde{c}_{ii'}^t = 1 - c_{ii'}^t$. $\tilde{Y}_{ii'}^{*l} = 1 - Y_{ii'}^{*l}$.

La maximisation précédente revient à (cf. Marcotorchino, Michaud [6]) :

$$\text{Min} \sum_{i,i'} \sum_{l=1}^{\nu} \sum_{t \in c_l} |c_{ii'}^t - Y_{ii'}^{*l}| \quad (2)$$

On détermine ensuite Y^{*l} la solution de ce problème d'optimisation par l'utilisation d'une méthode heuristique.

Les solutions ainsi trouvées sont des partitions de l'ensemble des individus en classes non imbriquées dont on n'a fixé a priori, ni la taille, ni le nombre.

En résolvant les ν modèles, (grâce aux nouvelles heuristiques développées dont le temps de calcul est négligeable), on obtient l'ensemble des variables qualitatives agrégées $J' = \{Y^{*1}, \dots, Y^{*l}, \dots, Y^{*\nu}\}$.

3.2.1 Validation de notre méthodologie dans le dernier cas.

Soit P_V la partition initiale obtenue à partir de toutes les variables à l'aide du critère de Condorcet, ce qui revient à maximiser le critère :

$$\sum_{i,i'} (c_{ii'}^t x_{ii'} + \tilde{c}_{ii'} \tilde{x}_{ii'}) \equiv \sum_{i,i'} \sum_{t=1}^m (c_{ii'}^t x_{ii'} + \tilde{c}_{ii'} \tilde{x}_{ii'}) \quad (3)$$

$$\Leftrightarrow \text{Min} \sum_{i,i'} \sum_{t=1}^m |c_{ii'}^t - x_{ii'}|.$$

De plus (3) peut s'écrire :

$$\sum_{i,i'} \left(\sum_{l=1}^{\nu} \left(\sum_{t \in c_l} c_{ii'}^t \right) x_{ii'} + \sum_{l=1}^{\nu} \left(\sum_{t \in c_l} \tilde{c}_{ii'}^t \right) \tilde{x}_{ii'} \right).$$

Ensuite, on pondère les variables Y^{*l} , ($l = 1, \dots, \nu$), en remplaçant Y^{*l} par $|c_l| Y^{*l}$ où $|c_l|$ est le cardinal de la classe c_l . Notons $R = \{|c_1| Y^{*1}, \dots, |c_l| Y^{*l}\}$ l'ensemble des variables pondérées. Afin d'obtenir la partition finale P_R , on maximise le critère :

$$\sum_{l=1}^{\nu} \sum_{i,i'} [(|c_l| Y_{ii'}^{*l}) x'_{ii'} + (|\tilde{c}_l| \tilde{Y}_{ii'}^{*l}) \tilde{x}'_{ii'}].$$

Pour valider notre méthodologie il faut montrer que l'écart entre la partition finale P_R , et celle obtenue sur la totalité des variables, notée P_V , est faible. Pour cela on va montrer que :

$$\sum_{i,i'} \sum_{l=1}^{\nu} \left| \sum_{t \in c_l} c_{ii'}^t - |c_l| Y_{ii'}^{*l} \right|$$

est bornée par la quantité (2) qui est minimisée par les $Y_{ii'}^{*l}$.

En effet

$$\sum_{i,i'} \sum_{l=1}^{\nu} \left| \sum_{t \in c_l} c_{ii'}^t - |c_l| Y_{ii'}^{*l} \right| \leq \sum_{i,i'} \sum_{l=1}^{\nu} \sum_{t \in c_l} |c_{ii'}^t - Y_{ii'}^{*l}|.$$

et $\sum_{i,i'} \sum_{l=1}^{\nu} \sum_{t \in c_l} |c_{ii'}^t - Y_{ii'}^{*l}|$ est minimale d'après la relation (2), donc $\sum_{i,i'} \sum_{l=1}^{\nu} \left| \sum_{t \in c_l} c_{ii'}^t - |c_l| Y_{ii'}^{*l} \right|$ est bornée par une valeur minimale qui concerne les Y^{*l} .

Remarque.

Pour vérifier que P_R est proche de P_V , il suffit de mesurer l'association entre les deux partitions en utilisant le critère de Rand.

4. Exemple illustratif

Afin d'illustrer le principe de l'algorithme, nous reprenons le classique exemple des «canidés» [8].

Les données du tableau (1.1) décrivent les caractéristiques de 27 races de chiens au moyen de variables qualitatives, les 6 premières considérées comme actives la septième «fonction» comme supplémentaire : ses trois modalités sont «compagnie», «chasse», «utilité».

TABLEAU 1.1

	Taille	Poids	veloc	Intel.	Affect.	Agre.	Fonction
1.Beauceron	3	2	3	3	2	2	3
2.Basset	1	1	1	1	1	2	2
3.B.Allemand	3	2	3	3	2	2	3
4.Boxer	2	2	2	2	2	2	1
5.Bull-Dog	1	1	1	2	2	1	1
6.Bull-Mastiff	3	3	1	3	1	2	3
7.Caniche	1	1	2	3	2	1	1
8.Chihuahua	1	1	1	1	2	1	1
9.Cocker	2	1	1	2	2	2	1
10.Colley	3	2	3	2	2	1	1
11.Dalmatien	2	2	2	2	2	1	1
12.Doberman	3	2	3	3	1	2	3
13.Dogue Allemand	3	3	3	1	1	2	3
14.Epagneul Breton	2	2	2	3	2	1	2
15.E.Français	3	2	2	2	1	1	2
16.Fox-Hound	3	2	3	1	1	2	2
17.Fox-Terrier	1	1	2	2	2	2	1
18.G.B.de Gascogne	3	2	2	1	1	2	2
19.Labrador	2	2	2	2	2	1	2
20.Levrier	3	2	3	1	1	1	2
21.Mastiff	3	3	1	1	1	2	3
22.Pekinois	1	1	1	1	2	1	1
23.Pointer	3	2	3	3	1	1	2
24.Saint-Bernard	3	3	1	2	1	2	3
25.Setter	3	2	3	2	1	1	2
26.Teckel	1	1	1	2	2	1	1
27.Terre-Neuve	3	3	1	2	1	1	3

où : La Taille a 3 modalités : 1 : petite, 2 : moyenne, 3 : grande.

Le Poids a 3 modalités : 1 : léger, 2 : moyen, 3 : lourd.

La Vitesse a 3 modalités : 1 : faible, 2 : moyenne, 3 : forte.

L'Intelligence a 3 modalités : 1 : médiocre, 2 : moyenne, 3 : forte.

L'Affection a 2 modalités : 1 : oui, 2 : non.

L'Aggressivité a 2 modalités : 1 : oui, 2 : non.

La classification des variables, en utilisant le critère :

$$H_{tt'}^1 = \Omega(V_t, V_{t'}) - \tilde{\Omega}(V_t, V_{t'})$$

où $\Omega(V_t, V_{t'})$ est le critère de Rand, est la suivante :

Classe 1 V_1 taille. V_2 Poids. V_3 Vitesse.**Classe 2** V_4 Intelligence.**Classe 3** V_5 Affection.**Classe 4** V_6 Agressivité.

On obtient quatre classes dont une contient trois variables et les autres contiennent chacune une seule. Deux cas sont à envisager :

1) Dans ce cas la sélection se fait en fonction de la contribution de chaque variable par rapport à chaque classe. En calculant ces contributions, on obtient pour la première classe : $C(V_1, C_1) = 3.70$, $C(V_2, C_1) = 4.02$ et $C(V_3, C_1) = 3.78$, on en déduit que V_2 représente la classe 1.

2) Dans le deuxième cas, on agrège les variables de la classe 1 en un seul élément Y^{*1} et pour les classes restantes la variable consensus pour chaque classe est la variable elle-même.

La variable consensus, Y^{*1} , qui agrège les éléments de la classe 1 est la suivante :

Classe 1	Classe 2	Classe 3	Classe 4
1 Beauceron.	2 Basset.	4 Boxer.	6 Bull-Mastiff.
3 Berger allemand.	5 Bull-Dog.	11 Dalmatien.	21 Mastiff.
10 Colley.	7 Caniche.	14 Epagneul Breton.	24 Saint-Bernard.
12 Doberman.	8 Chihuahua.	19 Labrador.	25 Setter.
13 Dogue allemand.	9 Cocker.		
15 Epagneul Français.	17 Fox-Terrier.		
16 Fox-Hound.	22 Pékinois.		
18 G.B. de Gasogne.	26 Teckel.		
20 Lévrier.			
23 Pointer.			
27 Terre Neuve.			

La partition initiale obtenue, notée P_V , de 27 races de chiens avec 6 variables actives est la suivante :

Classe 1	Classe 2	Classe 3	Classe 4
1 Beauceron.	2 Basset.	4 Boxer.	10 Colley.
3 Berger allemand.	5 Bull-Dog.	11 Dalmatien.	15 Epagneul Français.
6 Bull-Mastiff.	7 Caniche.	14 Epagneul Breton.	25 Setter.
12 Doberman.	8 Chihuahua.	19 Labrador.	27 Terre-Neuve.
13 Dogue Allemand.	9 Cocker.		
16 Fox-Hound.	17 Fox-Terrier.		
18 G.B. de Gasogne.	22 Pékinois.		
20 Lévrier.	26 Teckel.		
21 Mastiff.			
23 Pointer.			
24 Saint-Bernard.			

La partition finale P_R en utilisant les variables agrégées $3Y^{*1}$, Y^{*2} , Y^{*3} , Y^{*4} (3 représente le cardinal de la classe 1), est la suivante :

Classe 1	Classe 2	Classe 3	Classe 4
1 Beauceron.	2 Basset.	4 Boxer.	6 Bull-Mastiff.
3 Berger allemand.	5 Bull-Dog.	11 Dalmatien.	21 Mastiff.
10 Colley.	7 Caniche.	14 Epagneul Breton.	24 Saint-Bernard.
12 Doberman.	8 Chihuahua.	19 Labrador.	27 Terre-Neuve.
13 Dogue allemand.	9 Cocker.		
15 Epagneul Français.	17 Fox-Terrier.		
16 Fox-Hound.	22 Pékinois.		
18 G.B. de Gasogne.	26 Teckel.		
20 Lévrier.			
23 Pointer.			
25 Setter.			

On remarque que les classes 2 et 3 de la partition finale P_R , et la partition initiale P_V restent invariantes, par contre les classes 1 et 4 subissent un petit changement.

Pour vérifier que P_R est proche de P_V , on mesure l'association entre ces deux partitions en utilisant le critère de Rand.

On a $R(P_R, P_V) = 0.82$. Donc il y a une très forte association entre P_R et P_V . On peut conclure que P_R est une bonne approximation de P_V , faite par un ensemble

de variables de taille moindre, ce processus généralisé montre l'intérêt qui pourra en découler sur des ensembles (variables, objets) de tailles plus grandes.

5. Conclusion

L'introduction de nouvelles structures (par des critères d'associations connus : Rand, Ecart à l'indétermination, Ecart à l'indépendance, χ^2 , ϕ^2 , Belson, dérivé de Jordan) $R - \tilde{R}$, $I - \tilde{I}$, $E - \tilde{E}$, $B - \tilde{B}$, afin de classifier l'espace des variables qualitatives, nous a permis de donner des bornes ou seuils au-delà desquels on considère que les deux variables sont « ressemblantes ».

De plus cette classification peut nous servir pour la sélection des variables qualitatives. Un autre point important est la possibilité de tout expliciter en fonction du critère de Rand ou du χ^2 .

La démarche qui consiste à réduire le nombre de variables qualitatives en utilisant ce critère, aboutit à des résultats proche de la classification sur la totalité des variables.

Des simulations sur des grands fichiers sont l'objet de recherches actuelles.

Références

- [1] ABDALLAH H., (1996), Application de l'analyse relationnelle pour classifier descripteurs et modalités en mode discrimination, Thèse de l'Université Pierre et Marie Curie.
- [2] CRAMER H., (1946), Mathematical methods of statistics, Princeton university press, Princeton.
- [3] GOODMAN L., KRUSKAL W., (1979), Measures of association for cross classifications, Springer-Verlag, New York.
- [4] LERMAN I.C., (1987), «Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque : Application au problème de consensus en classification», R.S.A., 2, 39-60.
- [5] MARCOTORCHINO F., (1984), Utilisation des comparaisons par paires en statistiques des contingences (Partie 1 et 2). Etude du centre scientifique IBM France.
- [6] MARCOTORCHINO F., MICHAUD P., (1979), Optimisation en analyse ordinale des données, Masson, Paris.
- [7] MESSATFA H., (1988), «Unification de certains critères d'associations par linéarisation et normalisation», R.S.A. Vol 3.
- [8] SAPORTA G., (1990), Probabilités, Analyse des données et statistiques, Éditions Technip.