REVUE DE STATISTIQUE APPLIQUÉE

H. BENHADDA
F. MARCOTORCHINO

Introduction à la similarité régularisée en analyse relationnelle

Revue de statistique appliquée, tome 46, n° 1 (1998), p. 45-69
<http://www.numdam.org/item?id=RSA_1998__46_1_45_0>

Rev. Statistique Appliquée, 1998, **XLVI** (1), 45-69

INTRODUCTION À LA SIMILARITÉ RÉGULARISÉE EN ANALYSE RELATIONNELLE

H. Benhadda, F. Marcotorchino

Centre européen de Mathématiques Appliquées, ECAM-IBM, 68-76 Quai de la Rapée, 75592 Paris cedex 12

RÉSUMÉ

Cet article, dans la droite ligne des travaux faits sur l'analyse relationnelle depuis ces quinze dernières années, présente une nouvelle approche, dite régularisée, dans la façon de considérer la similarité (ou mesure d'affinité) entre les individus d'une population à classifier.

Le principe de cette régularisation est général, car applicable à des données dont les variables descriptives sont de type quelconque.

Nous nous restreindrons ici à l'application de ce principe dans le cas où les variables descriptives sont de type qualitatif.

Mots-clés : *Classification, analyse relationnelle, similarité régularisée, variables qualitatives.*

ABSTRACT

In this paper, we present a new approach called "regularised similarity", related to the works done on relational analysis within the last fifteen years.

This approach amounts to consider a similarity (or measure of association) between individuals of a population to be clustered in a different flexible way than usually done.

The principle of this regularisation is general and suitable to fit data derived from any type of variables.

We limit our paper to the application of the regularised similarity in the case where the variables are of categorical type.

Keywords : *Clustering, relational analysis, regularised similarity, categorical variables.*

1. Introduction

Classifier[1] un ensemble $\mathcal{I}$ de n individus décrits par un ensemble $\mathcal{V}$ de m variables, revient à partager cet ensemble en groupes homogènes, où à l'intérieur de chaque groupe les individus sont plus «similaires» entre eux, qu'ils ne le sont aux éléments des autres groupes.

[1] Dans le sens d'un partitionnement de cet ensemble.

Cette similarité, entre deux individus donnés de $\mathcal{I}$, se présente le plus souvent sous la forme d'un «indice de similarité» global[2] prenant ses valeurs dans l'intervalle [0,1], mais il est rarement fait référence de façon explicite à leur similarité «unitaire»[3] .

Or, que signifie «être similaire» lorsqu'on s'intéresse à une seule variable?

Cette approche, jouant sur la similarité unitaire, a été entrevue et étudiée par un certain nombre d'auteurs dont Anderberg [1], Goodall [9], et plus récemment Milioli [22].

Goodall, faisant l'hypothèse restrictive d'indépendance des variables, après avoir défini la similarité entre deux modalités d'une même variable V^k, a défini la similarité entre deux individus i et i' par rapport à cette variable comme le complément à 1 de la probabilité que deux individus pris au hasard aient, pour cette variable, deux modalités dont la similarité est plus grande ou égale à celle de la paire observée.

Anderberg quant à lui, partant de l'idée que des modalités rares d'une variable doivent avoir des poids plus importants que ceux des modalités plus courantes a défini deux indices de similarité comme fonctions inverses de la probabilité de concordance par rapport à une modalité donnée. Milioli, dans l'article cité ci-dessus, traitant le cas particulier où l'ensemble des variables est un questionnaire, a proposé de pondérer la concordance par rapport à une question par le nombre de modalités de cette question divisé par le nombre total de modalités de toutes les variables.

Partant de l'idée intuitive que plus le nombre de modalités d'une variable est grand, moins il est probable que deux individus pris au hasard dans $\mathcal{I}$ possèdent la même modalité de cette variable, et après avoir défini une similarité unitaire appelée «similarité unitaire logique», nous définirons deux nouveaux types de similarité dites «similarités régularisées».

Ces nouvelles similarités seront déduites de la similarité logique par pondération de cette dernière par des coefficients adéquats, dans le but de prendre en compte les disparités entre les nombres de modalités des variables d'une part et des effectifs de ces modalités d'autre part.

Nous noterons de façon générique s^k la similarité unitaire par rapport à une variable V^k et nous donnerons dans chaque cas son expression relationnelle et, quand cela est nécessaire, son expression vectorielle[4] .

La similarité «globale», que l'on notera s, sera définie comme la somme des similarités «unitaires» par rapport à chaque variable :

$$s = \sum_{k=1}^{m} s^k \tag{1}$$

Cette similarité s définie sur l'ensemble $\mathcal{I} \times \mathcal{I}$ et à valeurs dans l'ensemble des réels positifs $\mathbb{R}^+$ doit vérifier les propriétés suivantes :

---

[2] C'est-à-dire tenant compte des profils des deux individus par rapport à l'ensemble des variables.
[3] C'est-à-dire par rapport à une seule variable.
[4] Expression liée à la forme disjonctive complète de la variable V^k.

i) (La symétrie)

$$s_{ii'} = s_{i'i} \qquad \forall\, i, i' \in \mathcal{I}$$

ii) (L'auto-similarité maximale)

$$s_{ii} \geqslant s_{ii'} \qquad \forall\, i' \in \mathcal{I}$$

Une autre façon d'exprimer cette propriété est :

$$s_{ii'} \leqslant Min(s_{ii}, s_{i'i'}) \qquad \forall\, i, i' \in \mathcal{I}$$

A chaque mesure de similarité s sera associée une mesure de dissimilarité $\overline{s}$, qui sera définie comme suit :

Définition 1 *La dissimilarité globale $\overline{s}_{ii'}$ entre deux individus i et i', est égale au complément de leur similarité $s_{ii'}$ à la moyenne arithmétique de leur similarité propre s_{ii} et $s_{i'i'}$.*

$$\overline{s}_{ii'} = \frac{s_{ii} + s_{i'i'}}{2} - s_{ii'} \tag{2}$$

L'affectation des individus aux classes de la partition recherchée, sera fondée sur la vérification d'une condition simple et intuitive, qu'on appellera «règle de similarité positive», dont la définition est :

Définition 2 *On appellera «règle de similarité positive» entre deux individus i et i', la condition :*

$$s_{ii'} \quad \geqslant \quad \overline{s}_{ii'} \tag{3}$$

De façon formelle, on définira la similarité régularisée comme suit :

Définition 3 *Une similarité «unitaire» $s^{k'}$, par rapport à une variable donnée V^k, sera dite «régularisée» si elle est le produit d'une similarité initiale s^k par un coefficient α^k dépendant de la structure de cette variable et permettant de «lisser» l'influence de cette structure :*

$$s^{k'} \quad = \quad s^k \alpha^k$$

2. Rappels

Supposons que nous disposions de données relatives à un ensemble $\mathcal{I}$ de n individus décrits par un ensemble $\mathcal{V}$ de m variables homogènes de type qualitatif notées : $V^1, \ldots, V^k, \ldots, V^m$ chacune ayant respectivement : $p_1, \ldots, p_k, \ldots, p_m$ modalités. Notons p le nombre total de modalités toutes variables confondues :

$$p = \sum_{k=1}^{m} p_k$$

2.1. Le tableau disjonctif complet

Il est d'usage de mettre ces données sous la forme d'un tableau $\mathcal{K}$ de dimensions $(n \times p)$, appelé tableau disjonctif complet, de terme général k_{ij} tel que :

$$k_{ij} = \begin{cases} 1 & \text{si l'individu } i \text{ possède la modalité } j \\ 0 & \text{sinon} \end{cases} \tag{4}$$

Chaque marge colonne $k_{.j}$ ($j = 1, 2, \ldots, p$) du tableau $\mathcal{K}$ représente le nombre d'individus de l'ensemble $\mathcal{I}$ possédant la modalité j :

$$k_{.j} = \sum_{i=1}^{n} k_{ij} \tag{5}$$

Chaque marge ligne $k_{i.}$ ($i = 1, 2, \ldots, n$) de $\mathcal{K}$ représente le nombre de modalités possédées par l'individu i. Le tableau $\mathcal{K}$ a la propriété intéressante que toutes ses marges lignes sont égales[5] à m.

2.2. Les tableaux relationnels

Les tableaux relationnels sont basés sur le principe de comparaisons par paires. Ils croisent l'ensemble $\mathcal{I}$ avec lui même et sont donc de dimensions $(n \times n)$.

2.2.1. Le tableau individuel de Condorcet

Le tableau relationnel noté C^k, que l'on appelle «tableau de Condorcet» associé à la variable qualitative V^k, est de terme général $c^k_{ii'}$ tel que :

$$c^k_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ ont la même modalité de } V^k \\ 0 & \text{sinon} \end{cases} \tag{6}$$

Sous forme vectorielle $c^k_{ii'}$ est donnée par l'équation :

$$c^k_{ii'} = \sum_{j=1}^{p_k} k_{ij} k_{i'j} \tag{7}$$

[5] En effet, si les variables descriptives sont de type qualitatif et qu'il n'y a pas de données manquantes (ce que nous supposons ici), chaque individu possède une et une seule des modalités de chaque variable.

D'où l'on déduit que la somme des n^2 éléments du tableau relationnel C^k est :

$$c_{..}^k = \sum_{i=1}^{n}\sum_{i'=1}^{n} c_{ii'}^k \tag{8}$$

$$= \sum_{j=1}^{p_k} k_{.j}^2 \tag{9}$$

Remarque :

Comme V^k est une variable qualitative, elle correspond à une relation d'équivalence. La propriété de réflexivité se traduit pour le tableau C^k par la relation :

$$c_{ii}^k = 1 \qquad \forall\, i \in \mathcal{I} \tag{10}$$

Au tableau C^k est associé le tableau complémentaire $\overline{C}^k$ de terme général $\overline{c}_{ii'}^k$ tel que :

$$\overline{c}_{ii'}^k = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ n'ont pas la même modalité de } V^k \\ 0 & \text{sinon} \end{cases} \tag{11}$$

S'il n'y a pas de données manquantes, ce que nous supposerons dans la suite de l'exposé, on a la propriété relationnelle suivante :

$$c_{ii'}^k + \overline{c}_{ii'}^k = 1 \qquad \forall\, i, i' \in \mathcal{I} \tag{12}$$

2.2.2. Le tableau collectif de Condorcet

Le tableau relationnel collectif[6] noté C, que l'on appelle simplement «tableau de Condorcet» a pour terme général $c_{ii'}$ tel que :

$$c_{ii'} = \sum_{k=1}^{m} c_{ii'}^k$$

$c_{ii'}$ est donc égal au nombre de variables pour lesquelles i et i' partagent les mêmes modalités.

La somme des n^2 éléments du tableau relationnel collectif C est :

$$c_{..} = \sum_{k=1}^{m} c_{..}^k \tag{13}$$

[6] C'est-à-dire relatif à l'ensemble $\mathcal{V}$ de toutes les variables

Le tableau relationnel complémentaire $\overline{C}$, associé au tableau C, a pour terme général $\overline{c}_{ii'}$ tel que :

$$\begin{aligned}\overline{c}_{ii'} &= \sum_{k=1}^{m} \overline{c}_{ii'}^{k} \\ &= m - c_{ii'} \qquad (\textit{cf.}\ \text{eq. } (12))\end{aligned}$$

$\overline{c}_{ii'}$ est égal au nombre de variables pour lesquelles i et i' ne partagent pas les mêmes modalités.

on a donc la propriété relationnelle globale :

$$c_{ii'} + \overline{c}_{ii'} = m \qquad (14)$$

Remarque :

De l'égalité (10) on déduit, par sommation sur les m variables, que :

$$c_{ii} = m \qquad (15)$$

3. La similarité logique

3.1. Similarité unitaire

La similarité unitaire logique entre les deux individus i et i' par rapport à la variable V^k correspondra, par définition, au codage relationnel C^k de cette variable, soit :

$$s_{ii'}^{k} = c_{ii'}^{k} \qquad (16)$$

En utilisant la relation (7), l'écriture vectorielle de cette similarité est :

$$s_{ii'}^{k} = \sum_{j=1}^{p_k} k_{ij} k_{i'j} \qquad (17)$$

Des définitions (2) et (17), on déduit directement les expressions[7] relationnelles et vectorielles de la dissimilarité $\overline{s}_{ii'}^{k}$, soit :

$$\begin{aligned}\overline{s}_{ii'}^{k} &= \overline{c}_{ii'}^{k} && \text{(expression relationnelle)} \\ &= \frac{1}{2}\sum_{j=1}^{p_k}(k_{ij} - k_{i'j})^2 && \text{(expression vectorielle)}\end{aligned}$$

[7] *cf.* Marcotorchino [16].

3.2. Similarité globale

La similarité globale logique $s_{ii'}$ entre i et i', par rapport à l'ensemble des m variables, est par définition égale au nombre de variables pour lesquelles i et i' partagent les mêmes modalités (*cf.* eq. (1)) :

$$s_{ii'} = c_{ii'}$$

Cette similarité est nulle, lorsque les profils des deux individus sont totalement distincts; elle est maximum, de valeur m (nombre total de variables), lorsque les profils des deux individus sont identiques.

$$0 \leqslant s_{ii'} \leqslant m$$

De la même façon, la dissimilarité globale est :

$$\overline{s}_{ii'} = \overline{c}_{ii'}$$

3.3. Règle d'affectation

La règle de similarité positive, dans ce cas, revient à :

$$c_{ii'} \geqslant \overline{c}_{ii'} \tag{18}$$

Ce qui correspond, d'après l'égalité (14), à la règle à la «majorité» de Condorcet (*cf.* Michaud [21]) :

$$c_{ii'} \geqslant \frac{m}{2}$$

Interprétation :

Deux individus i et i' seront *a priori*[8] affectés à la même classe de la partition recherchée dès lors que le nombre de variables qu'ils partagent est supérieur ou égal à la moitié du nombre total des variables de départ.

En conclusion, la similarité globale associée à la similarité unitaire logique, est égale au nombre de concordances entre les deux profils relatifs à i et i'.

On retrouve donc, sans modification, l'approche traditionnelle du comptage des occurrences de concordance de l'approche Condorcet.

[8] En effet, il ne faut pas oublier que le problème général de classification se heurte à la non transitivité par paires (ou effet Condorcet) des similarités.

4. La similarité statistique

Cette similarité, qu'on peut qualifier de «présence-rareté», tient compte non seulement de la modalité partagée par les deux individus, mais aussi du nombre d'individus possédant cette modalité dans $\mathcal{I}$.

Plus ces deux individus sont «rares» à partager la même modalité de V^k, plus ils sont semblables et *vice-versa*.

4.1. Similarité unitaire

Si $c^k_{i.}$ est le nombre d'individus ayant la même modalité que l'individu i pour la variable V^k :

$$c^k_{i.} = \sum_{i'=1}^{n} c^k_{ii'} \tag{19}$$

alors, la similarité statistique entre les deux individus i et i', est telle que :

$$s^k_{ii'} = \begin{cases} \dfrac{1}{c^k_{i.}} & \text{si } i \text{ et } i' \text{ ont la même modalité de } V^k \\ 0 & \text{sinon} \end{cases} \tag{20}$$

elle s'écrit aussi, de façon plus synthétique, sous la forme :

$$s^k_{ii'} = \frac{c^k_{ii'}}{c^k_{i.}} \tag{21}$$

Une autre écriture relationnelle plus symétrique de $s^k_{ii'}$ est donnée par l'expression suivante :

$$s^k_{ii'} = \frac{c^k_{ii'}}{2}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right) \tag{22}$$

Remarque :

D'une part, cette similarité, de façon implicite, fait jouer un rôle au nombre de modalités[9] de la variable V^k à travers la valeur de $c^k_{i.}$; et d'autre part, la similarité d'un individu avec lui même n'est plus une valeur constante égale à 1 comme dans le cas précédent.

En conclusion, $s^k_{ii'}$ dans ce cas est bien un indice de «présence-rareté», tenant compte par son numérateur de la présence logique d'une modalité commune à i et i' et par son dénominateur de la rareté de cette modalité.

[9] En effet, plus le nombre de modalités de V^k est grand, moins les effectifs des modalités ont la chance d'être forts.

Il est facile de montrer, dans ce cas, que la propriété d'auto-similarité maximale est bien vérifiée. En effet,

$$\begin{aligned} s^k_{ii'} &= \frac{c^k_{ii'}}{c^k_{i.}} \\ &\leqslant \frac{1}{c^k_{i.}} \quad (\text{car } c^k_{ii'} \leqslant 1) \\ &= s^k_{ii} \qquad \forall\, i' \in \mathcal{I} \end{aligned}$$

En utilisant la définition (2), on montre que :

$$\overline{s}^k_{ii'} = \frac{\overline{c}^k_{ii'}}{2}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right) \tag{23}$$

4.2. Similarité globale

Dans ce cas, la similarité globale entre i et i' n'est autre que le coefficient de Condorcet pondéré (*cf.* Marcotorchino [16]) :

$$s_{ii'} = \sum_{k=1}^{m} \frac{c^k_{ii'}}{c^k_{i.}} \tag{24}$$

Qui peut s'écrire aussi sous une forme plus symétrique :

$$s_{ii'} = \frac{1}{2}\sum_{k=1}^{m} c^k_{ii'}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right) \tag{25}$$

La dissimilarité globale entre i et i' est, quant à elle, donnée par :

$$\overline{s}_{ii'} = \frac{1}{2}\sum_{k=1}^{m} \overline{c}^k_{ii'}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right) \tag{26}$$

Si l'on pose :

$\mathcal{A}_{ii'}$ = l'ensemble des variables par rapport auxquelles i et i' sont en concordance :

$$\mathcal{A}_{ii'} = \{k \in \mathcal{V} / c^k_{ii'} = 1\}$$

$\overline{\mathcal{A}}_{ii'}$ = l'ensemble des variables par rapport auxquelles i et i' ne sont pas en concordance :

$$\overline{\mathcal{A}}_{ii'} = \{k \in \mathcal{V} / \overline{c}^k_{ii'} = 1\}$$

$h^k_{ii'}$ = la moyenne harmonique de l'effectif de la modalité possédée par i et de l'effectif de la modalité possédée par i' par rapport à la variable V^k :

$$\frac{2}{h^k_{ii'}} = \frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}$$

$h_{ii'}$ = la moyenne harmonique des effectifs des modalités partagées par i et i' sur l'ensemble des variables :

$$\begin{aligned}\frac{c_{ii'}}{h_{ii'}} &= \sum_{k\in\mathcal{A}_{ii'}} \frac{1}{c^k_{i.}}\\ &= \frac{1}{2}\sum_{k\in\mathcal{A}_{ii'}}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right)\end{aligned}$$

en particulier, on a :

$$\begin{aligned}\frac{c_{ii}}{h_{ii}} &= \frac{m}{h_{ii}} \qquad (\textit{cf.}\ \text{eq. } (15))\\ &= \sum_{k=1}^{m} \frac{1}{c^k_{i.}}\end{aligned}$$

$\overline{h}_{ii'}$ = la moyenne harmonique des effectifs des modalités non partagées par i et i' sur l'ensemble des variables :

$$\frac{\overline{c}_{ii'}}{\overline{h}_{ii'}} = \frac{1}{2}\sum_{k\in\overline{\mathcal{A}}_{ii'}}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right)$$

$\widetilde{h}_{ii'}$ = la moyenne harmonique des moyennes harmoniques des effectifs des modalités possédées par i et des effectifs des modalités possédées par i' sur l'ensemble des variables :

$$\frac{2}{\widetilde{h}_{ii'}} = \frac{1}{h_{ii}} + \frac{1}{h_{i'i'}} \tag{27}$$

En utilisant la formule (24) et la définition (2), il est facile de montrer que :

$$s_{ii'} = \frac{c_{ii'}}{h_{ii'}} \tag{28}$$

$$\overline{s}_{ii'} = \frac{\overline{c}_{ii'}}{\overline{h}_{ii'}} \tag{29}$$

Des relations (25), (26) et de la propriété relationnelle (12) on déduit :

$$s_{ii'} + \overline{s}_{ii'} = \frac{1}{2}\sum_{k=1}^{m}\left(\frac{1}{c^k_{i.}} + \frac{1}{c^k_{i'.}}\right) \tag{30}$$

$$= \frac{m}{2}\left(\frac{1}{h_{ii}} + \frac{1}{h_{i'i'}}\right) \tag{31}$$

$$= \frac{m}{\widetilde{h}_{ii'}} \qquad (\textit{cf.}\ \text{eq. } (27)) \tag{32}$$

et d'après les équations (28) et (29), on obtient la relation :

$$\frac{c_{ii'}}{h_{ii'}} + \frac{\overline{c}_{ii'}}{\overline{h}_{ii'}} = \frac{m}{\widetilde{h}_{ii'}} \tag{33}$$

Cette formule est à comparer à la formule (14) correspondant au cas d'une similarité logique.

Remarque :

Des relations (15) et (28), on déduit que la similarité statistique globale d'un individu avec lui même est :

$$s_{ii} = \frac{m}{h_{ii}}$$

La propriété d'auto-similarité maximale implique donc :

$$\frac{c_{ii'}}{h_{ii'}} \leqslant Min\left(\frac{m}{h_{ii}}, \frac{m}{h_{i'i'}}\right)$$

soit

$$\frac{c_{ii'}}{m} \leqslant \frac{h_{ii'}}{Max(h_{ii}, h_{i'i'})}$$

ou ce qui est équivalent :

$$\frac{c_{ii'}}{Max(c_{ii}, c_{i'i'})} \leqslant \frac{h_{ii'}}{Max(h_{ii}, h_{i'i'})} \tag{34}$$

4.3. Règle d'affectation

La règle de similarité positive s'écrit sous la forme :

$$\frac{c_{ii'}}{h_{ii'}} \geqslant \frac{\overline{c}_{ii'}}{\overline{h}_{ii'}} \tag{35}$$

Soit dans ce cas :

$$\frac{c_{ii'}}{h_{ii'}} \geqslant \frac{m}{\widetilde{h}_{ii'}} - \frac{c_{ii'}}{h_{ii'}} \qquad \text{(d'après (33))}$$

d'où :

$$c_{ii'} \geqslant \frac{m}{2} \frac{h_{ii'}}{\widetilde{h}_{ii'}} \tag{36}$$

De ce fait, suivant la valeur du rapport $\frac{h_{ii'}}{\widetilde{h}_{ii'}}$, cette règle sera plus ou moins généreuse[10] que la règle à la majorité de Condorcet.

En effet,

- si $\widetilde{h}_{ii'} > h_{ii'}$, la règle est plus généreuse que celle de Condorcet.
- si $\widetilde{h}_{ii'} < h_{ii'}$, la règle est moins généreuse que celle de Condorcet.

5. La similarité probabiliste

Cette similarité, que l'on pourrait qualifier de «présence-difficulté», tient compte de la probabilité pour deux objets pris au hasard de partager une variable donnée.

Contrairement à la similarité logique où «être similaire» pour deux objets par rapport à une variable qualitative V^k revient pour eux à partager la même modalité de cette variable, indépendamment du nombre de modalités qu'elle possède et des effectifs de ces dernières, cette nouvelle similarité, fait intervenir un aspect probabiliste qui tiendra compte de la difficulté intrinsèque pour deux objets quelconques, pris au hasard dans la population d'origine, d'être en concordance par rapport à V^k.

En effet, il semble intuitif de considérer qu'il est d'autant plus difficile à deux objets d'être similaires par rapport à une variable[11] que le nombre de modalités de celle-ci est grand.

Le principe de la «similarité unitaire» est alors de tenir compte par une pondération dite «régularisée» de la difficulté relative, à deux individus d'être similaires, par rapport à une variable donnée.

En d'autres termes, cette pondération cherchera à équilibrer le rôle du nombre de modalités dans l'occurrence d'une «concordance» pour ne pas favoriser les variables à peu de modalités par rapport aux variables à beaucoup de modalités.

---

10 Le terme «généreux» ici veut dire que le seuil de comparaison de la similarité régularisée est plus faible que celui de la similarité logique.

11 C'est-à-dire partager la même modalité.

5.1. Similarité unitaire

Si l'on note δ_j la probabilité pour un individu donné i de posséder la modalité j de la variable V^k alors, du fait de l'indépendance[12] des individus, la probabilité que deux individus i et i' possèdent une modalité j est égale au produit des probabilités qu'a chaque individu de posséder cette modalité.

Comme par définition les modalités sont exclusives, la concordance de i et i' par rapport à V^k ne pouvant se réaliser que sur une et une seule des p_k modalités de V^k, on en déduit que la probabilité de concordance π_k entre i et i' par rapport à V^k est :

$$\pi_k = \sum_{j=1}^{p_k} \delta_j^2 \tag{37}$$

On définira donc la similarité unitaire probabiliste (ou de difficulté) entre i et i' par rapport à la variable V^k par :

$$s_{ii'}^k = c_{ii'}^k(1 - \pi_k) \tag{38}$$

5.1.1. Probabilité théorique

Si l'on considère que les p_k modalités de la variable V^k sont équiprobables, alors :

$$\delta_j = \frac{1}{p_k} \tag{39}$$

d'où, d'après (37) :

$$\pi_k = \sum_{j=1}^{p_k} \frac{1}{p_k^2} \tag{40}$$

$$= \frac{1}{p_k} \tag{41}$$

5.1.2. Probabilité empirique

On estime ici la probabilité de présence δ_j de la modalité j par la proportion d'individus possédant cette modalité :

$$\delta_j = \frac{k_{.j}}{n} \tag{42}$$

[12] En effet, la possession d'une modalité par un individu i n'influe en rien la possession ou la non possession de cette modalité par un individu i'.

d'où :

$$\pi_k = \sum_{j=1}^{p_k} \frac{k_{.j}^2}{n^2} \tag{43}$$

$$= \frac{c_{..}^k}{n^2} \qquad (\textit{cf.}\ \text{eq. } (9)) \tag{44}$$

5.2. Similarité globale probabiliste

La similarité globale probabiliste entre deux individus i et i' est :

$$s_{ii'} = \sum_{k=1}^{m} c_{ii'}^k (1 - \pi_k) \tag{45}$$

En utilisant la définition (2), on trouve que la dissimilarité globale, entre i et i' est donnée par la formule suivante :

$$\overline{s}_{ii'} = \sum_{k=1}^{m} \overline{c}_{ii'}^k (1 - \pi_k) \tag{46}$$

En utilisant la propriété relationnelle (12), on déduit que :

$$s_{ii'} + \overline{s}_{ii'} = \sum_{k=1}^{m} (1 - \pi_k) \tag{47}$$

$$= m - \sum_{k=1}^{m} \pi_k \tag{48}$$

5.2.1. Similarité globale probabiliste (Cas théorique)

Dans ce cas :

$$s_{ii'} = \sum_{k=1}^{m} c_{ii'}^k \left(1 - \frac{1}{p_k}\right)$$

$$= c_{ii'} - \sum_{k=1}^{m} \frac{c_{ii'}^k}{p_k}$$

En utilisant les définitions des ensembles $\mathcal{A}_{ii'}$ et $\overline{\mathcal{A}}_{ii'}$ données en [4.2.] on obtient :

$$s_{ii'} = c_{ii'} - \sum_{k \in \mathcal{A}_{ii'}} \frac{1}{p_k} \tag{49}$$

Si l'on pose :

$p_{ii'}$ = la moyenne harmonique du nombre de modalités des variables partagées par i et i' :

$$\frac{c_{ii'}}{p_{ii'}} = \sum_{k \in \mathcal{A}_{ii'}} \frac{1}{p_k}$$

$\overline{p}_{ii'}$ = la moyenne harmonique du nombre de modalités des variables non partagées par i et i' :

$$\frac{\overline{c}_{ii'}}{\overline{p}_{ii'}} = \sum_{k \in \overline{\mathcal{A}}_{ii'}} \frac{1}{p_k}$$

$\mathcal{P}$ = la moyenne harmonique des nombres de modalités de toutes les variables :

$$\frac{m}{\mathcal{P}} = \sum_{k=1}^{m} \frac{1}{p_k}$$

Comme un individu est en concordance avec lui même sur l'ensemble des variables, on a en particulier l'égalité :

$$p_{ii} = \mathcal{P} \qquad \forall\, i \in \mathcal{I} \tag{50}$$

L'équation de la similarité globale probabiliste (49) devient donc :

$$s_{ii'} = c_{ii'} \left(1 - \frac{1}{p_{ii'}}\right) \tag{51}$$

Des relations (15) et (50), on déduit que la similarité globale probabiliste d'un individu i avec lui même est une constante :

$$s_{ii} = m \left(1 - \frac{1}{\mathcal{P}}\right) \qquad \forall\, i \in \mathcal{I} \tag{52}$$

Remarque :

On posera par convention :

$$p_{ii'} = 1 \qquad (\text{si } \mathcal{A}_{ii'} = \varnothing)$$

Comme $|\mathcal{A}_{ii'}| + |\overline{\mathcal{A}}_{ii'}| = m$, on a :

$$\begin{aligned}\sum_{k\in\mathcal{A}_{ii'}} \frac{1}{p_k} + \sum_{k\in\overline{\mathcal{A}}_{ii'}} \frac{1}{p_k} &= \sum_{k=1}^{m} \frac{1}{p_k} \\ &= \frac{m}{\mathcal{P}}\end{aligned}$$

Soit :

$$\frac{c_{ii'}}{p_{ii'}} + \frac{\overline{c}_{ii'}}{\overline{p}_{ii'}} = \frac{m}{\mathcal{P}} \tag{53}$$

Remarque :

Cette formule est à rapprocher des formules (14) et (33).

En utilisant la définition (2) et d'après les relations (51) et (52), on déduit que la dissimilarité globale probabiliste entre i et i' est donnée par :

$$\begin{aligned}\overline{s}_{ii'} &= s_{ii} - s_{ii'} \\ &= m\left(1 - \frac{1}{\mathcal{P}}\right) - c_{ii'}\left(1 - \frac{1}{p_{ii'}}\right) \\ &= (m - c_{ii'}) - \left(\frac{m}{\mathcal{P}} - \frac{c_{ii'}}{p_{ii'}}\right) \\ &= \overline{c}_{ii'} - \frac{\overline{c}_{ii'}}{\overline{p}_{ii'}} \qquad (\textit{cf.}\ \text{eqs. } (14)(53))\end{aligned}$$

Soit :

$$\overline{s}_{ii'} = \overline{c}_{ii'}\left(1 - \frac{1}{\overline{p}_{ii'}}\right) \tag{54}$$

5.2.2. Règle d'affectation

La règle de la similarité positive se traduit alors, dans le cas de la probabilité de concordance théorique, par :

$$c_{ii'}\left(1 - \frac{1}{p_{ii'}}\right) \geqslant \overline{c}_{ii'}\left(1 - \frac{1}{\overline{p}_{ii'}}\right) \qquad \forall\, i, i' \tag{55}$$

ou bien par :

$$c_{ii'}\left(1 - \frac{1}{p_{ii'}}\right) \geqslant \frac{m}{2}\left(1 - \frac{1}{\mathcal{P}}\right) \tag{56}$$

Remarque :

Si la règle de similarité positive est vérifiée pour deux individus i et i', on peut facilement montrer que la relation (55) implique les inégalités suivantes :

$$c_{ii'}\left(1-\frac{1}{p_{ii'}}\right) \geqslant \frac{m}{2}\left(1-\frac{1}{\mathcal{P}}\right) \geqslant \overline{c}_{ii'}\left(1-\frac{1}{\overline{p}_{ii'}}\right)$$

En résumé, dans le cas d'une probabilité de concordance théorique, la règle de similarité positive,

a) sera plus généreuse que la règle de Condorcet si :

$$\frac{1}{p_{ii'}} < \frac{1}{\mathcal{P}}$$

soit

$$\mathcal{P} < p_{ii'}$$

b) sera équivalente à la règle de Condorcet si :

$$\mathcal{P} = p_{ii'}$$

c) sera moins généreuse que la règle de Condorcet si :

$$\mathcal{P} > p_{ii'}$$

5.2.3. Similarité globale (Cas empirique)

Dans ce cas, la similarité globale est :

$$\begin{aligned} s_{ii'} &= \sum_{k=1}^{m} c_{ii'}^{k}\left(1-\frac{c_{..}^{k}}{n^2}\right) \\ &= c_{ii'} - \sum_{k=1}^{m} c_{ii'}^{k}\frac{c_{..}^{k}}{n^2} \\ &= c_{ii'} - \sum_{k\in\mathcal{A}_{ii'}} \frac{c_{..}^{k}}{n^2} \end{aligned}$$

et la dissimilarité globale est :

$$\overline{s}_{ii'} = \overline{c}_{ii'} - \sum_{k\in\overline{\mathcal{A}}_{ii'}} \frac{c_{..}^{k}}{n^2}$$

5.2.4. Règle d'affectation

La règle de similarité positive $s_{ii'} \geqslant \overline{s}_{ii'}$ donne ici :

$$c_{ii'} - \sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{n^2} \geqslant \overline{c}_{ii'} - \sum_{k \in \overline{\mathcal{A}}_{ii'}} \frac{c_{..}^k}{n^2} \tag{57}$$

et comme :

$$\sum_{k \in \mathcal{A}_{ii'}} c_{..}^k + \sum_{k \in \overline{\mathcal{A}}_{ii'}} c_{..}^k = c_{..}$$

La relation (57) devient, en utilisant la propriété relationnelle (14) :

$$2c_{ii'} \geqslant m + \left(2 \sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{n^2} - \frac{c_{..}}{n^2}\right)$$

soit :

$$c_{ii'} \geqslant \frac{m}{2} + \frac{c_{..}}{n^2} \left(\sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{c_{..}} - \frac{1}{2}\right)$$

En résumé, dans le cas d'une probabilité de concordance empirique, la règle de similarité positive sera :

a) plus généreuse que la règle de Condorcet si :

$$\sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{c_{..}} < \frac{1}{2}$$

b) équivalente à la règle de Condorcet si :

$$\sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{c_{..}} = \frac{1}{2}$$

c) moins généreuse que la règle de Condorcet si :

$$\sum_{k \in \mathcal{A}_{ii'}} \frac{c_{..}^k}{c_{..}} > \frac{1}{2}$$

6. Conclusion

Le problème (P) de la classification de la population $\mathcal{I}$ des n individus dans l'approche relationnelle consiste à chercher une partition de $\mathcal{I}$, que nous représenterons par une matrice booléenne X de dimensions $(n \times n)$ et de terme général $x_{ii'}$ tel que :

$$x_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ sont dans la même classe de } X \\ 0 & \text{sinon} \end{cases} \tag{58}$$

que nous appellerons aussi X, par abus de langage, et qui maximisera le critère $\mathcal{C}(X)$ suivant :

$$\mathcal{C}(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} (s_{ii'}^{\alpha} x_{ii'}^{\beta} + \overline{s}_{ii'}^{\alpha} \overline{x}_{ii'}^{\beta}) \tag{59}$$

Les indices α et β représentent le type de similarité choisi pour la matrice «données» et pour la matrice «solution».

Ces indices peuvent être :

l pour la similarité logique.

r pour la similarité de rareté (statistique).

p pour la similarité probabiliste.

En résumé, on raisonnera désormais en termes de choix de similarité entre individu et non plus en termes de critère de classification, même si tout critère est implicitement relié à un type de similarité.

Grâce à la notion de régularisation, il dépendra désormais de l'utilisateur de fixer l'approche similarité qu'il veut introduire dans le critère qu'il utilise, compte tenu du but qu'il poursuit.

Une fois la similarité par rapport aux variables descriptives choisie, il pourra lui associer différents types de similarité[13] pour la partition inconnue X.

A titre d'exemple (*cf.* Marcotorchino et Benhadda [17]) :

- Si $\alpha = \beta = l$, c'est-à-dire qu'on cherche à associer une similarité logique pour les données avec une similarité logique pour la partition inconnue, on retrouve le critère de Condorcet (*cf.* Michaud [21]).
- Si $\alpha = r$ et $\beta = l$, c'est-à-dire qu'on cherche à associer une similarité statistique pour les données avec une similarité logique pour la partition inconnue, on retrouve le critère de Condorcet pondéré (*cf.* Marcotorchino [16]).

[13] Logique, Statistique ou Probabiliste.

Sous forme plus synthétique, le problème (P) de la classification relationnelle est :

$$(P) \quad \begin{cases} \max\limits_{X} \mathcal{C}(X) & & \\ x_{ii} = 1 & \forall\, i \in \mathcal{I} & \text{(réflexivité)} \\ x_{ii'} - x_{i'i} = 0 & \forall\, (i,i') \in \mathcal{I}^2 & \text{(symétrie)} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leqslant 1 & \forall\, (i,i',i'') \in \mathcal{I}^3 & \text{(transitivité)} \\ x_{ii'} \in \{0,1\} & \forall\, (i,i') \in \mathcal{I}^2 & \text{(binarité)} \end{cases}$$

Remarques :

- Il est à noter que dans l'approche relationnelle, il n'y a pas de fixation *a priori* du nombre de classes de la partition solution. Ce nombre est donné de façon automatique par la méthode.
- Pour une étude de la similarité régularisée dans le cas où les variables sont de type continu, se reporter à [17].
- L'étude de la similarité régularisée dans le cas où les variables sont de type fréquentiel et binaire «présence-absence» sera traitée dans la thèse de H. Benhadda [4].

En conclusion, l'approche «similarité régularisée», outre le fait qu'elle permet de «corriger» l'influence de la disparité de la structure modalitaire des variables, permet également de raisonner non plus en termes de critère de classification, mais en termes «d'association» de types de similarité (similarité type des données versus similarité type de la solution).

7. Exemple d'application

Nous donnons ici un exemple d'application de la similarité régularisée sur des données concernant un ensemble $\mathcal{I}$ de vingt félins décrits par onze variables. Cet exemple est extrait de l'article de Marcotorchino et Michaud [19].

Les félins sont répartis comme suit :

Félins d'Afrique :

Lion, Léopard, Guépard, Serval, Caracal, Chat Cafer et Chat Dore.

Félins d'Eurasie :

Tigre, Once, Léopard, Panthère Nébuleuse, Lynx, Chat de Borneo, Chat Malais, Chat du Bengale et Chat de Chine.

Félins d'Amérique :

Jaguar, Puma, Chat Tigrin, Chat des Andes et Chat Marbré.

Ces félins sont décrits par les onze variables qualitatives suivantes :

V^1= «Grimpe aux arbres» (1 :Oui, 0 :Non)

V^2= «Chasse à courre ou à l'affût» (1 :Oui, 0 :Non)

V^3= «Comportement prédateur» (1 Diurne, 2 :Diurne ou nocturne, 3 :Nocturne)

V^4= «Canines développées» (1 :Oui, 0 :Non)

V^5= «Fourrure» (1 :Poils longs, 0 :Poils ras)

V^6= «Longueur du corps» (1 :Petite, 2 :Moyenne, 3 :Grande)

V^7= «Oreilles» (1 :Rondes ou arrondies, 2 :En pointe)

V^8= «Poids» (1 :faible, 2 :Moyen, 3 :Fort)

V^9= «Queue» (1 :Petite, 2 :Moyenne, 3 :Longue)

V^{10}= «Taille au garot» (1 :Petite, 2 :Moyenne, 3 :Grande)

V^{11}= «Aspect du pelage» (1 :Sans taches, 2 :Tacheté, 3 :Rayé, 4 :Marbré)

Les données sont résumées dans la tableau suivant :

	V^1	V^2	V^3	V^4	V^5	V^6	V^7	V^8	V^9	V^{10}	V^{11}
Lion	0	1	1	1	0	3	1	3	2	3	1
Tigre	0	0	3	1	0	3	1	3	2	3	3
Jaguar	1	0	2	1	0	2	1	3	1	3	2
Léopard	1	0	3	1	0	2	1	3	2	3	2
Puma	1	0	2	1	0	2	1	3	3	2	1
Lynx	1	0	2	1	1	2	2	2	1	2	2
Chine	1	0	2	0	0	1	2	1	1	1	1
Serval	1	1	1	0	0	2	2	2	1	2	2
Caracal	1	1	2	0	0	1	2	2	1	2	1
Guépard	0	1	1	0	0	2	1	2	3	3	2
Borneo	1	0	3	0	0	1	1	1	2	1	1
Malais	1	0	3	0	1	1	1	1	1	1	1
Bengale	1	0	3	0	0	1	1	1	2	1	2
Tigrin	1	0	3	0	0	1	1	1	2	1	2
Dore	1	0	3	0	0	1	1	1	2	1	1
Cafer	1	1	3	0	0	1	1	1	2	1	3
Marbré	1	0	3	0	0	1	1	1	3	1	4
Andes	1	0	3	0	1	2	1	1	2	1	2
Once	1	0	1	1	1	2	1	2	3	2	2
Nébul	1	0	3	1	0	2	1	2	3	2	4

7.1. Les différentes partitions obtenues

Ces partitions sont toutes legèrement différentes les unes des autres, ce qui montre l'influence des coefficients de régularisation.

Partition logique L'utilisation de la similarité logique nous donne une partition en quatre classes :

Classe 1 :

Borneo, Dore , Malais, Bengale, Tigrin, Cafer, Marbré, Chine, Andes.

Classe 2 :

Lynx, Once, Puma, Jaguar, Nébul, Léopard.

Classe 3 :

Serval, Guépard, Caracal.

Classe 4 :

Lion, Tigre.

Partition probabiliste (cas théorique) L'utilisation de la similarité probabiliste théorique nous donne une partition en cinq classes :

Classe 1 :

Borneo, Dore , Malais, Bengale, Tigrin, Cafer, Marbré, Chine, Andes.

Classe 2 :

Lynx, Once, Puma, Jaguar, Nébul, Léopard.

Classe 3 :

Serval, Caracal.

Classe 4 :

Lion, Tigre.

Classe 5 :

Guépard.

Partition probabiliste (cas empirique) L'utilisation de la similarité probabiliste empirique nous donne une partition en cinq classes :

Classe 1 :

Borneo, Dore , Malais, Bengale, Tigrin, Cafer, Marbré, Chine, Andes.

Classe 2 :

Lion, Tigre, Jaguar, Léopard.

Classe 3 :

Lynx, Once, Puma, Nébul.

Classe 4 :

Serval, Caracal.

Classe 5 :

Guépard.

Partition statistique L'utilisation de la similarité statistique nous donne une partition en six classes :

Classe 1 :

Borneo, Dore , Malais, Bengale, Tigrin, Cafer, Marbré, Andes.

Classe 2 :

Serval, Caracal, Lynx, Chine.

Classe 3 :

Jaguar, Puma, Léopard.

Classe 4 :

Lion, Tigre.

Classe 5 :

Once, Nébul.

Classe 6 :

Guépard.

7.2. Essai d'interprétation

Nous remarquons qu'il existe dans l'ensemble $\mathcal{I}$, relativement aux variables retenues, cinq classes stables de félins :

$\mathcal{C}_1$ = {Borneo, Dore , Malais, Bengale, Tigrin, Cafer, Marbré, Andes}.

$\mathcal{C}_2$ = {Serval, Caracal}.

$\mathcal{C}_3$ = {Lion, Tigre}.

$\mathcal{C}_4$ = {Jaguar, Léopard}.

$\mathcal{C}_5$ = {Once, Nébul}.

Partition logique : Cette partition étant obtenue par un simple comptage du nombre de variables partagées par les individus, a attiré le Guépard vers la classe stable $\mathcal{C}_2$. C'est le Serval qui a été à l'origine de cette affectation car il partage sept variables sur onze avec le Guépard. Cette partition est la seule à ne pas isoler le Guépard.

Partition Probabiliste (cas théorique) : Cette partition a séparé le Guépard de la classe stable $\mathcal{C}_2$ car le Guépard et le Caracal ne partagent pas sept variables sur onze dont les plus significatives, en termes de poids dans le calcul des similarités, sont les variables V^3, V^6, V^9, V^{10} à trois modalités chacune et la variable V^{11} à quatre modalités.

Partition Probabiliste (cas empirique) : Cette partition a réuni les classes stables $\mathcal{C}_3$ et $\mathcal{C}_4$ car le Tigre et le Léopard partagent sept variables sur onze dont les plus significatives, en termes de poids dans le calcul des similarités, sont les variables V^1, V^2 et V^3.

C'est d'ailleurs d'un point de vue de la taxinomie zoologique, celle qui est la plus significative (*cf.* Dorst et Dandelot [8]).

Partition statistique : Cette partition pondérant fortement les concordances sur des modalités à faible effectif, a attiré le Puma vers la classe stable $\mathcal{C}_4$ car le Puma partage huit modalités avec le Jaguar dont les modalités deux de V^3 et trois de V^8 qui ne sont possédées que par cinq individus chacune dans $\mathcal{I}$.

Elle a de même réuni le Chine et le Lynx car ils partagent les modalités deux de V^3, deux de V^7 et une de V^9. Ces deux individus ont été attirés vers la classe stable $\mathcal{C}_2$ car le Chine et le caracal partagent huit modalités dont les plus significatives, en termes de poids dans le calcul des similarités, sont la modalité deux de V^3, la modalité deux de V^7 et la modalité une de V^9.

Références bibliographiques

[1] ANDERBERG M.R., *Cluster analysis for application*, Academic Press, New York, 1973.

[2] BAULIEU F.B., A classification of presence/absence based dissimilarity coefficients, *Journal of Classification*, vol. 6, 1989, pp. 233-246.

[3] BEDECARRAX C., Classification en analyse relationnelle : La quadri-décomposition et ses application, *Thèse de l'Université Paris VI*, 1989.

[4] BENHADDA H., La similarité régularisée et ses applications en classification automatique, *Thèse de l'Université de Paris VI*, 1998.

[5] BURNABY T.P., On a method for character weighting similarly coefficient, employing the concept of information, *Math. Geol.*, vol. 2, n° 1, 1970, pp. 25-38.

[6] CAILLIEZ F., PAGES J.-P., *Introduction à l'analyse des données*, Paris, Smash, 1976.

[7] DECAESTECKER C., Apprentissage en classification conceptuelle incrémentale, *Thèse de l'Université Libre de Bruxelles (Faculté des Sciences)*, 1992.

[8] DORST J., DANDELOT P., *Guide des grands mammifère d'Afrique*, Neuchâtel, Delachaux et Niestlé, 1972.

[9] GOODALL D.W., A new similarity index based on probability, *Biometrics*, n° 22, 1966, pp. 882-907.

[10] GORDON A.D., Constructing dissimilarity measures, *Journal of Classification*, vol. 7, 1990, pp. 257-269.

[11] GOWER J.-C., A note on Burnaby's character weighted similarity coefficient, *Math. Geol.*, vol. 2, n° 1, 1970, pp. 39-45.

[12] GOWER J.-C., A general coefficient of similarity and some of its properties, *Biometrics*, vol. 27, 1971, pp. 857-872.

[13] JOLY S., et LE CALVÉ G., Similarity functions, *Lecture Notes in Statistics, (In Van Cutsem, B. ed) Springer-Verlag*, vol. 93, 1994, pp. 67-86.

[14] LERMAN I.C., Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque, application au problème du consensus en classification (1), *Revue de Statistique Appliquée*, vol. 35, n° 2, 1987, pp. 39-60.

[15] MARCOTORCHINO F., La classification automatique aujourd'hui : bref aperçu historique applicatif et calculatoire, *Publications Scientifiques et Techniques d'IBM France*, n° 2, Novembre 1991, pp. 35-93.

[16] MARCOTORCHINO F., L'analyse factorielle-relationnelle : parties I et II, *Etude du CEMAP, IBM France*, vol. Etude MAP-03, décembre 1991.

[17] MARCOTORCHINO F., BENHADDA H., Introduction à la théorie générale de la similarité régularisée, *Etude du CEMAP, IBM France*, vol. Etude MAP-011, Juillet 1996.

[18] MARCOTORCHINO F., EL AYOUBI N., Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association, *Revue de Statistique Appliquée*, vol. 39, n° 2, 1991, pp. 25-46.

[19] MARCOTORCHINO F., MICHAUD P., Agrégation de similarités en classification automatique, *Revue de Statistique Appliquée*, vol. 30, n° 2, 1982.

[20] MICHAUD P., Agrégation à la majorité I : hommage à Condorcet, *Etude du Centre Scientifique IBM France*, vol. F051, 1982.

[21] MICHAUD P., Hommage à Condorcet (version intégrale pour le bicentenaire de l'essai de Condorcet); *Etude du Centre Scientifique IBM France*, vol. F.094, 1985.

[22] MILIOLI M.A., A similarity index between multiple choice categorical answers in surveys, *In : Distancia'92*, pp. 127-130, Rennes, 22-26 Juin 1992.

[23] SAPORTA G., *Probabilité analyse des données et statistique*, Technip, 1990.

[24] SNIJDERS T. A.B. *et al.*, Distribution of some similarity coefficients for dyadic binary data in the case of associated attributes, *Journal of Classification* n° 7, 1990, pp. 5-31.