REVUE DE STATISTIQUE APPLIQUÉE

J. M. CALLOUD

Deux algorithmes de classification sur des tableaux de contingence ou de correspondance

Revue de statistique appliquée, tome 18, n° 4 (1970), p. 41-45

<http://www.numdam.org/item?id=RSA_1970__18_4_41_0>

© Société française de statistique, 1970, 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/

DEUX ALGORITHMES DE CLASSIFICATION SUR DES TABLEAUX DE CONTINGENCE OU DE CORRESPONDANCE

J. M. CALLOUD

Centre de Calcul - CHU Pitié Salpétrière - Paris

Nous allons essayer de présenter au cours de ce bref exposé deux récents algorithmes de classification hiérarchique, propres au traitement des tableaux de contingence ou des tableaux de correspondance.

Mais "rendons à César ce qui est à César", cette communication doit beaucoup du cours de 3ème Cycle, sur la classification automatique, professé par Monsieur BENZECRI (1) en 1968-69 à la Faculté des Sciences de Paris, cours auquel l'on pourra se reporter avec beaucoup de profit pour plus de détails et de rigueur.

Nous commencerons par un bref rappel de la théorie de l'information, puis nous définirons ce que nous entendons par information mutuelle entre variables avant d'aborder les algorithmes proprement dits.

I - RAPPEL DE LA THEORIE DE L'INFORMATION

L'information apportée par l'assertion : "il s'est produit tel événement a" que l'on notera H(a) vérifie les axiomes suivants :

1/ H(a) est une fonction réelle positive et décroissante de la seule probabilité p(a)

2/ si deux événements sont indépendants, l'information apportée par la connaissance des deux événements est égale à la somme des informations apportées par chacun d'eux.

3/ (axiome d'échelle) un événement qui a une chance sur deux d'être réalisé apporte une information unité. De ces trois axiomes on déduit que :

$$\forall P \in [0,1] \qquad f(P) = -\log_2(P)$$

D'autre part, soit $i \in I$, i aléatoire de la loi P_I (ou $P_I = [p_i \mid i \in I]$ le système des probabilités de chacun des éléments de l'ensemble I).

Il est naturel d'associer à l'assertion "la loi de I est P_I" une quantité d'information H qui sera l'espérance mathématique de l'information apportée par l'assertion élémentaire "la probabilité d'avoir i est p_i".

---

(1) Laboratoire de Statistique Mathématique Tour 45 - 2ème Etage - Faculté des Sciences de Paris - Publications : Inf. Tab. "Théorie de l'Information et Classification".

$$H(P_I) = -\ \Sigma\{p_i\ \mathrm{Log}_2\ p_i \mid i \in I\}$$

II - INFORMATION MUTUELLE ENTRE VARIABLES

Soit (i , j) une paire, élément aléatoire de l'ensemble produit $I \times J$, soient P_{IJ}, P_I , P_J les lois de la paire (i , j) et de chacun de ses membres. Où

$$P_{IJ} = \{P_{i,j} \mid i \in I,\ J \in J\}$$

$$\forall i \in I,\ \forall j \in J\ \ p(i , j) \geqslant 0 \text{ et } \Sigma\{p_{i,j} \mid i \in I,\ j \in J\} = 1$$

$$P_I = \{p_i \mid i \in I\} \text{ avec } \quad p_i = \Sigma\{p_{i,j} \mid j \in J\}$$

alors

$$H(P_{IJ}) \leqslant H(P_I) + H(P_J)$$

Soient I les arrondissements de Paris et J les catégories socio-professionnelles : alors connaître l'arrondissement où habite une personne donne une présomption sur la profession de celle-ci ; par suite l'information $H(P_{IJ})$ apportée par la connaissance de l'arrondissement et de la profession est inférieure ou égale à la somme des informations apportées par la connaissance de l'arrondissement seul $H(P_I)$ et de la profession seule $H(P_J)$.

$H(P_{IJ}) = H(P_I) + H(P_J)$ lorsque les deux ensembles sont indépendants. C'est-à-dire $\forall i \in I,\ \forall i \in J\ p_{ij} = p_i * p_j$. On peut alors poser $H(p_{IJ}) - H(p_I) = H(J/I)$ information conditionnelle apportée par j quand i est connu.

De même $H(P_J) - H(J/I) = H(p_I) + H(p_J) - H(p_{IJ})$ information sur J apportée par I que l'on notera $H(p_{IJ} ; p_I * p_J)$.

L'information apportée par I sur J a une expression symétrique en I et J, aussi peut-on l'appeler information mutuelle entre I et J et la considérer comme une mesure de la dépendance (de l'interaction) entre I et J).

III - INFORMATION MUTUELLE ET DISTANCE DU χ^2

L'information mutuelle est une quantité qui s'annule quand p_{IJ} est une loi produit $p_I * p_J$, on calcule de même, sous le nom de χ^2, une quantité critère qui sert à comparer p_{IJ} et $p_I * p_J$.

On a pour expression du carré de la distance entre ces deux lois dans la métrique du χ^2 de centre $p_I * p_J$

$$\| P_{IJ} - p_I * p_J \| = \sum\{(p_{ij} - p_i p_j)\ / p_i p_j / i \in I,\ j \in J\}$$

$$= \sum_{ij}\{p_i p_j (p_{ij}/p_i p_j - 1)^2\}$$

Or nous avons d'autre part

$$H(P_{IJ} ; P_I * P_J) = \sum_{ij} \{p_{ij} \ \mathrm{Log}_2 (p_{ij}/p_i p_j)\}$$

$$= \sum_{ij} \left\{ p_i p_j * \frac{p_{ij}}{p_i p_j} \ \mathrm{Log}_2 \left(\frac{p_{ij}}{p_i p_j} \right) \right\}$$

Ces deux expressions (χ^2 et information mutuelle) ne diffèrent qu'en ce que $x^2 - 1$ a été remplacé par $x \log_2 x$. Ces deux fonctions, comme l'a montré Monsieur BENZECRI, sont osculatrices à tangente horizontale au point $x = 1$, on ne doit donc pas s'étonner que les algorithmes qui en dérivent conduisent pratiquement aux mêmes résultats. Remarquons que l'algorithme basé sur la formule de χ^2 ne nécessitant pas le calcul d'un logarithme est nettement plus rapide.

IV - LES ALGORITHMES DE CLASSIFICATION

Soit un tableau de contingence rectangulaire $I \times J$, $k(i,j)$ sera le nombre de personnes dans l'arrondissement i exerçant la profession j. De ce tableau on passe à une loi de probabilité sur $I \times J$ par le calcul des fréquences $\forall i \in I, \ \forall j \in J \ \ p_{ij} = k(i,j) / \sum_{ij} k(i,j)$.

La dépendance entre i et j est la seule information dont on dispose pour organiser les ensembles I et J soit H (PIJ ; PI P3) ou $||P_{IJ} - P_I * P_J||^2$

Par suite, la ressemblance entre i et i' se mesurera par la ressemblance entre leur profils respectifs sur J. C'est-à-dire entre P_J^i et $P_J^{i'}$ $\left(P_J^i = \{P_j^i\} \text{ et } P_J^i = \frac{p_{ij}}{p_i}\right)$.

Soit Q une partition de I, cette partition sera d'autant plus conforme au principe de proximité qu'à l'intérieur d'une classe les éléments se ressembleront le plus, c'est-à-dire que P_J^i variera peu.

Or connaitre la classe à laquelle appartient un individu apporte moins d'information que la connaissance de l'individu lui-même, un certain nombre d'attributs propres à l'individu n'ayant pas été pris en compte pour permettre cette partition, par suite

$$H(P_{QJ} ; P_Q * P_J) \leqslant H(P_{IJ} ; P_I P_J)$$

L'égalité n'ayant lieu que si à l'intérieur de chaque classe les individus sont identiques.

Les algorithmes de construction ascendante que nous proposons, se traduisant par la recherche pas à pas de partitions de I de moins en moins fines, notre critère sera de rendre minimale l'écart, à chaque pas de la construction, par rapport à la situation idéale que nous venons d'envisager.

Notons après Monsieur BENZECRI

$$\mathrm{Lien}_1 (I , J) = H(P_{IJ} ; P_I * P_J)$$

$$\text{et } \mathrm{Lien}_2 (I , J) = ||P_{IJ} - P_I * P_J||^2$$

La quantité critère que nous aurons à minimiser sera donc à chaque pas

$$\Lambda = \text{Lien}_{1 \text{ ou } 2}(I , J) - \text{Lien}_{1 \text{ ou } 2}(Q , J)$$

Supposons construite une hiérarchie A_{h-1} - notons Som A_{h-1} l'ensemble des sommets de cette hiérarchie (c'est-à-dire l'ensemble des classes q de la partition de I associée à la hiérarchie).

Le passage de A_{h-1} à A_h se fera par la réunion de deux sommets de A_{h-1}, s et s' tels que la quantité critère Λ correspondante soit minimales parmi toutes les quantités associées à toutes les réunions possibles.

On montre que lorsque l'on utilise la distance dite du χ^2 (Lien_2), la distance entre les sommets prend alors la forme suivante :

$$d(s , s') = \frac{p_s \circ p'_s}{p_s + p'_s} \quad || s - s' ||^2$$

coefficient de masse. — distance entre centre de gravité des classes s et s'

(p_s = poids de la classe s)

L'effet du coefficient de masse est qu'à distance égale $||s - s'||^2$, deux classes sont agrégées d'autant plus bas dans la hierarchie qu'elles sont plus légères.

Nous venons de présenter brièvement, trop peut-être, deux algorithmes que l'on doit, si l'on suit Monsieur TOMASSONNE, ranger dans la catégorie des algorithmes basés sur un concept de masse.

RESUME

Ce bref exposé présente deux algorithmes de classifications hiérarchiques propres au traitement des tableaux de contingence ou de correspondance (tableaux I * J avec I : ensemble des individus et J : ensemble des mesures effectuées sur ces individus).

A la base de ces deux algorithmes se trouve une mesure de la dépendance (ou du lien) entre les ensembles I et J. Cette mesure dérive dans un cas de la théorie de l'information et dans l'autre de celle du χ^2.

Ces algorithmes, étant donné le rôle important joué par la notion de masse (masse des individus et masse des classes), séparent au mieux les zones à forte concentration en poids. Ils sont, croyons-nous, à rapprocher fructueusement des méthodes d'analyse factorielle (en particulier de l'analyse des correspondances de Monsieur BENZECRI).

DISCUSSION

Question de Monsieur LELLOUCH

Les deux distances définies par l'orateur définissent-elles des ultramétriques ?

Réponse : les algorithmes présentés conduisant à une hiérarchie totale indicée (ou arbre ou dendrogramme) définissent sur l'ensemble des éléments à classer une structure ultramétrique. En effet on peut montrer (cf Benzécri (1)) qu'il y a équivalence entre hiérarchie totale indicée de parties et structure ultramétrique.

Question de Monsieur VALLERON

Les p(j) qui interviennent dans la formule de la distance du χ^2 que vous avez donnée ne sont pas, en pratique, des probabilités. Ce sont en réalité les fréquences (choisies par l'expérimentateur) des objets à classer. On sait bien que ce choix influence le résultat final de la classification. Mais il me semble que dans le cas de la distance du χ^2 ce phénomène est encore amplifié. Qu'en pensez-vous ?

Réponse : nous avons précisé que ces algorithmes étaient propres au traitement des tableaux de contingence et des tableaux de correspondance. Les p(i , j), p(i), p(j) ne sont pas, et vous avez raison de le signaler, en pratique, des probabilités mais des fréquences. Ces fréquences ne sont pas choisies par l'expérimentateur, mais dérivent directement du tableau initial. Le passage aux fréquences nous permet de calculer les distances entre les profils des individus à classer, nous disposons alors d'une matrice de distances. Nous avons le choix entre deux possibilités : soit affecter chaque individu d'un poids égal à sa fréquence p(i) soit l'affecter d'un poids unité. Le poids d'une classe étant la somme des poids des éléments qui la composent. La première possibilité fait en effet jouer un rôle important aux fréquences p(i) (cf d(s , s') = ...) ; en revanche la deuxième établit une complète égalité entre les individus, le poids des classes devenant égal au nombre d'individus qui la composent. On doit choisir l'une ou l'autre possibilité suivant le type de problème à traiter.

Question de Monsieur LOCQUIN

Avec la formule $d(s , s') = ... || ||^2$ vous avez trouvé un phénomène fondamental celui de l'influence de la néguentropie sur la circulation de l'information. Celle-ci "courbe" celle là comme la gravité "courbe" la lumière en relativité.

Réponse : Je ne peux que vous laisser l'entière responsabilité de votre affirmation.