JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

J. LORIGNY

Application de la théorie de l'information à l'optimisation des questionnaires

Journal de la société statistique de Paris, tome 116 (1975), p. 212-218

<http://www.numdam.org/item?id=JSFS_1975__116__212_0>

APPLICATION DE LA THÉORIE DE L'INFORMATION A L'OPTIMISATION DES QUESTIONNAIRES

At first, the author mentions as a reminder the main property of the information quantity applied to questionnaires. He then describes the extension of this property to entry-exit graphs. At last he concludes by quoting some applications of those theoretical results.

Der Verfasser ruft zuerst die hauptsächlichsten Charakteristika der auf den Fragebogen gestellten Fragen in Erinnerung. Er beschreibt anschliessend die Charakteristika fur die graphischen Darstellungen : Eingang-Ausgang. Er beendet seine Studie indem er einige Anwendungen diser theoretischen Resultate gibt.

I — Introduction

La mesure de la quantité d'information d'une distribution de probabilités, à partir de la formule d'entropie de Shannon $H = -\Sigma\, p_i \log p_i$ est utilisée dans certains domaines du calcul statistique.

Par exemple, la distance entre deux distributions, au sens de Kullback, est une quantité d'information qui peut servir de critère pour la recherche d'un tableau croisé à marges données qui soit aussi voisin que possible d'un tableau donné. On peut aussi utiliser la quantité d'information de Shannon pour prévoir quel sera le nombre minimum de questions de base b donnée, qu'il faudra poser pour séparer complètement un ensemble d'issues (ou « réponses ») supposé muni d'une distribution de probabilités, que cette distribution de probabilités soit connue implicitement ou estimée dans une étape précédente.

Dans ce dernier ordre d'idées, s'est développée depuis une dizaine d'années la « théorie des questionnaires » et s'est formé un groupe de recherche au C. N. R. S. consacré aux Structures de l'information, dirigé par Claude-François Picard.

Au cours de ces dernières années, un des nombreux aspects de la recherche dans ce domaine, a consisté à étendre certains résultats déjà classiques pour les questionnaires arborescents, aux questionnaires latticiels et finalement à des graphes très généraux que nous appelons graphes d'entrée-sortie.

Ces travaux mériteraient d'être mieux connus des statisticiens et notre propos est une sorte d'invitation à découvrir (ou redécouvrir) cette approche « informationnelle ». Pour une étude plus précise des concepts employés, on peut recommander comme ouvrage de base, celui de C.-F. Picard : *Graphes et Questionnaires*, Gauthier-Villars, Paris, 1972.

Nous rappellerons d'abord brièvement la principale propriété de la quantité d'information de Shannon dans les questionnaires arborescents, puis dans les questionnaires latticiels. Ensuite, nous exposerons d'une façon succincte l'extension de cette propriété aux graphes d'entrée-sortie. Enfin, nous conclurons en évoquant quelques applications de ces différents résultats théoriques.

II — Questionnaires arborescents

On appelle questionnaire arborescent, un ensemble arborescent de questions emboîtées permettant finalement de différencier les éléments d'un ensemble. On appelle « issues », ou « réponses », ou « sommets terminaux », ou « sommets de sortie », ces éléments à différencier. La première question posée est appelée « racine » ou « sommet initial » ou « sommet d'entrée » de l'arborescence. On appelle « base » d'une question, le demi-degré extérieur du sommet c'est-à-dire le nombre d'arcs sortant de cette question. La base mesure par conséquent le pouvoir de différenciation de la question. Lorsque toutes les questions sont de même base, on dit que le questionnaire est homogène.

L'ensemble des sommets terminaux $S = (s_i : i = 1 \text{ à } I)$ est supposé muni d'une distribution de probabilités $p(s_i)$. En remontant le questionnaire arborescent, on affecte, de proche en proche, à chaque question intermédiaire ainsi qu'à la racine, une valuation égale à la somme des valuations des questions qui lui succèdent dans le questionnaire. La valuation obtenue pour la racine, est évidemment égale à 1. Le questionnaire étant arborescent on peut considérer que les valuations de questions ainsi définies sont aussi des valuations d'arcs puisque à toute question sauf la racine correspond un arc entrant et inversement à tout arc correspond une question extrémité.

Enfin, il existe une bijection entre l'ensemble des sommets terminaux et l'ensemble des chemins $(c_i : i = 1 \text{ à } I)$ joignant la racine aux sommets terminaux donc « traversant » le questionnaire. On peut donc munir l'ensemble de ces chemins de la distribution de probabilité donnée pour les sommets terminaux : $p(c_i) = p(s_i) : i = 1 \text{ à } I$.

Par ailleurs, on définit pour ces chemins soit une longueur égale au nombre des arcs du chemin, soit une fonction de coût, égale à la somme des « coûts » donnés aux arcs empruntés, ou encore à la somme des « coûts » des questions traversées, ce qui revient au même pour les questionnaires arborescents.

La première propriété remarquable exprime que la longueur moyenne de cheminement à travers un questionnaire arborescent est égale à la somme des valuations des questions (racine et sommets intermédiaires) :

$$L = \sum_{X-S} p(x)$$

en appelant longueur de cheminement L, la quantité : $L = \sum_{I} l(c_i)\ p(c_i)$. (1)

On peut alors se servir de cette propriété, après avoir défini auparavant l'entropie ou quantité d'information de Shannon d'une question x par la quantité :

$$H(x) = \sum_{z \in Q^1(x)} \frac{p(z)}{p(x)} \log \frac{p(x)}{p(z)}$$

$Q^1(x)$ représentant l'ensemble des questions succèdant à la question x.

On en déduit : $\sum_{X} p(x)\ H(x) = -\sum_{I} p(s_i) \log p(s_i)$ (2)

Par ailleurs, on sait que l'entropie se trouve toujours située entre deux valeurs extrêmes caractéristiques :

$H(x) \geqslant 0$ hétérogénéité absolue : 1 seul arc sortant de la question x.

$H(x) \leqslant \log m(x)$ homogénéité absolue : m arcs sortant de x avec des valuations équiparties.

Si le questionnaire arborescent est homogène et de base b, on en déduit la seconde relation cherchée entre la longueur de cheminement (moyen) à travers le questionnaire et l'entropie « terminale » ou « de sortie » du questionnaire :

$$L \geqslant -\sum_I p(s_i) \log_b p(s_i) \tag{3}$$

Cette relation établit notamment l'existence de questionnaires de longueur de cheminement minimum. Ces questionnaires sont appelés L-optimaux.

Notons que le raisonnement est encore vrai pour les questionnaires hétérogènes, à condition de préciser que b représente cette fois-ci la plus grande des bases du questionnaire.

La relation (1) est utilisée pour la recherche de questionnaires L-optimaux, c'est-à-dire de longueur minimale et de bases données aboutissant à un ensemble donné de réponses munies de probabilités données. Plusieurs méthodes sont employées, par exemple, en appliquant des règles de substitution d'arc et d'échange de sous-arborescences entre deux questions. Les algorithmes les plus connus sont ceux de Shannon-Fano et de Huffmann.

La relation (2) exprime, avec des définitions précises couramment admises, que l'information traitée par l'ensemble des questions est égale à l'information transmise par le questionnaire.

Pour imager cette notion de longueur de cheminement, on pourrait dire que la relation (3) relie « l'épaisseur » d'un questionnaire, ou « temps de traversée », ou « inertie » du questionnaire à l'entropie terminale.

Cette relation est en général une inégalité stricte, même pour un questionnaire L-optimal et l'égalité n'a lieu que si toutes les questions sont de bases b exactement, avec des probabilités sortantes équiparties, c'est-à-dire si la valuation de toute question de rang r est égale à b^{-r}. On appelle affaiblissement d'un questionnaire, la différence entre la longueur de cheminement minimale et l'entropie de sortie.

On peut donner une interprétation informationnelle des algorithmes de Shannon-Fano et de Huffmann. Ainsi, la méthode de Shannon-Fano consiste à réaliser des partitions emboîtées de plus en plus fines de l'ensemble des réponses en cherchant à chaque stade l'équipartition maximum, donc en maximisant l'information traitée depuis la racine jusqu'aux réponses.

La méthode de Huffmann consiste, elle, à regrouper des sommets en partant des réponses et en remontant vers la racine, en cherchant à minimiser à chaque question créée, l'apport d'information.

III — Questionnaires latticiels

Les résultats précédents ont été d'abord généralisés aux questionnaires latticiels. La difficulté vient de ce qu'il n'y a plus bijection entre l'ensemble des sommets terminaux et l'ensemble des chemins allant de la racine à ces sommets terminaux. Les chemins sont plus nombreux que les réponses.

On appelle hypothèse de proportionnalité des flux, l'hypothèse suivant laquelle le choix de l'issue d'une question dans la base de sortie de cette question est indépendant de l'arc par lequel le chemin est entré dans la question. L'intérêt de cette hypothèse tient au fait que l'information traitée par le questionnaire latticiel atteint alors sa valeur maximale.

Les résultats du cas arborescent sont généralisables au moins partiellement au cas latticiel. Les relations (1) et (3) sont encore vraies. L'égalité (2) se transforme en inégalité, c'est-à-dire que l'information traitée devient supérieure à l'information transmise.

IV — Graphes d'entrée-sortie

Nous avons généralisé enfin les propriétés mathématiques précédentes à une catégorie très large de graphes avec circuits, les graphes d'entrée-sortie, ainsi définis :

On considère trois sous-ensembles de sommets dans ces graphes.

— les sommets d'entrée, pour lesquels le flux entrant est inférieur au flux sortant;

— les sommets d'articulation, pour lesquels le flux entrant est égal au flux sortant;

— les sommets de sortie, pour lesquels le flux entrant est supérieur au flux sortant.

On fait une hypothèse de connexité, à savoir que tout sommet du graphe est situé sur au moins un chemin descendant d'un sommet d'entrée et sur au moins un chemin ascendant à un sommet de sortie.

Les chemins d'entrée-sortie, c'est-à-dire allant d'un sommet d'entrée à un sommet de sortie, soit au bout d'un nombre fini d'arcs, soit par une infinité d'arcs à cause des boucles ou circuits possibles, jouent, comme dans les paragraphes précédents, un rôle particulier. La même difficulté que précédemment se présente pour l'interprétation probabiliste des valuations de sommets, par rapport à celles choisies pour les chemins d'entrée-sortie. En pratique, on fait encore l'hypothèse de proportionnalité des flux sur l'arborescence compatible, et on choisit donc comme valuation d'un chemin $[x_1\, x_2\, x_3 \ldots]$, où $x_1\, x_2\, x_3 \ldots$ représensentent les occurences de sommets rencontrés, la quantité :

$$\frac{\varphi(x_1\, x_2)}{\varphi(x_1)} \cdot \frac{\varphi(x_2\, x_3)}{\varphi(x_2)}$$

De cette valuation de chemin, on déduit une seconde valuation des chemins d'entrée-sortie dont on démontre qu'elle constitue une distribution de probabilités.

Notons que l'arborescence compatible est infinie. D'ailleurs, le calcul à la limite donne, au passage, des évaluations intéressantes pour majorer les résidus.

A ces graphes d'entrée-sortie, on généralise la relation (1) qui s'écrit maintenant sous la forme :

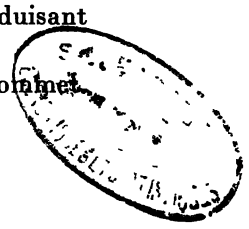
$$L = \frac{1}{\varphi} \sum_X \varphi(x) - 1$$

où $\varphi(x)$ est la valuation de sommet, X l'ensemble des sommets du graphe et φ le flux traversant le graphe, c'est-à-dire la somme des différences de flux pour l'ensemble des sommets d'entrée (ou des sommets de sortie).

A la relation (2) on peut faire correspondre une nouvelle relation (2′) en introduisant un nouveau concept informationnel :

La différence d'entropie entre les deux faces, amont et aval, d'un même sommet

$$\delta h(x) = H^{+1}(x) - H^{-1}(x)$$



La relation (2') s'obtient par simple sommation après avoir décomposé les expressions du second membre en leurs éléments simples :

$$\sum_X n(x)\ \delta h(x) = H_s - H_e = \Delta H(G) \qquad (2')$$

où H_e et H_s représentent respectivement l'entropie des flux entrant et sortant du graphe G.

Enfin, on peut généraliser la relation (3) précédente :

$$L \geqslant \Delta H_{BS}\ (G) - 1 \qquad (3')$$

pour le cas des graphes diffusants c'est-à-dire où l'entropie de sortie est plus grande que l'entropie d'entrée.

$$L \geqslant -\Delta H_{BE}(G) - 1$$

pour le cas des graphes concentrants c'est-à-dire où ΔH est négatif.

où BS (respectivement BE) représente le plus grand demi-degré extérieur (resp. intérieur) du graphe.

On établit que la longueur minimum de traversée (moyenne) d'un graphe d'entrée-sortie de spécifications externes données s'obtient quand il ne comporte ni boucle ni circuit, donc dans le cas du questionnaire latticiel précédemment étudié.

Notons que la borne donnée par la formule 3' n'est en général pas atteinte, même à l'optimum. Il faut pour cela que le latticiel soit une forêt d'arborescences, c'est-à-dire comporte un seul arc entrant par question, pour prendre le cas des graphes diffusants, et que pour chaque arborescence, la valuation des sommets de sortie de rang r soit b^{-r} comme indiqué précédemment dans le cas du questionnaire arborescent.

Enfin, une dernière généralisation concerne le cas des graphes avec coût :

D'abord, en partant d'une fonction de coût $c(\gamma)$ donnée sur l'ensemble Γ des arcs de G, on obtient la formule

$$C = \frac{1}{\nu} \sum_\Gamma c(\gamma)\ \nu(\gamma)$$

où la longueur de cheminement L est ici remplacée par le coût moyen C des chemins d'entrée-sortie (le coût d'un chemin vaut la somme des coûts de ses arcs) pondérés par la même distribution de probabilités que précédemment.

Ensuite, en partant d'une fonction de coût $c(x)$ donnée sur l'ensemble X des sommets de G, on obtient la formule :

$$C = \frac{1}{\nu} \sum_X c(x)\ \nu(x)$$

où C est le coût moyen des chemins d'entrée-sortie (le coût d'un chemin étant cette fois-ci la somme des coûts des sommets traversés) pondérés toujours par la même probabilité.

V — Conclusion

La théorie des questionnaires, a des applications pratiques dans tous les problèmes de cheminement dans un système de partitions d'un ensemble avec un objectif de longueur minimale ou de coût minimal en respectant une précision suffisante. La théorie des pseudo-

questionnaires (M. Terrenoire), qui en constitue une extension au cas aléatoire, a donné lieu en particulier à des applications avancées dans le domaine du diagnostic médical.

Montrons sur un exemple très simple l'application de la théorie de l'information à l'optimisation d'un questionnaire au sens large : considérons le problème de l'économie d'un code-identifiant.

Soit un fichier statistique de base destiné à une procédure permanente de sélections aléatoires, par exemple :

— un fichier de références bibliographiques servant de base aux réponses à la demande en documentation automatique;

— ou un fichier d'échantillon maître de logements servant de base de sondage à des enquêtes auprès des ménages;

— ou un fichier de données communales, servant de base à des interrogations à la demande;

— ou un fichier d'entreprises servant de base à des enquêtes auprès des entreprises,

— etc.

Chaque enregistrement d'individu est composé d'une zone-identifiant et d'une zone-données. Considérons l'opération de sélection sur la zone-identifiant.

On peut, au moins théoriquement, décomposer cette sélection en une arborescence de tests portant chacun sur un caractère de la zone-identifiant. En général, on utilise pour les fichiers statistiques des identifiants de longueur fixe. Alors, le temps de sélection d'un individu ayant un identifiant donné est constant quel que soit l'individu.

Mais supposons maintenant que le fichier statistique soit un fichier de base permanente, c'est-à-dire interrogé fréquemment et d'une façon très aléatoire avec de forts écarts entre les individus les plus fréquemment appelés et les individus les moins fréquemment appelés (c'est le cas des exemples donnés plus haut). Alors, il devient intéressant de remplacer l'identifiant de longueur fixe par un identifiant de longueur variable qui soit très court pour les individus très fréquents, très long pour les individus très rares de façon à minimiser le temps moyen de sélection, c'est-à-dire le nombre moyen de tests de caractères pour isoler un identifiant donné pondéré par la fréquence d'appel de l'individu.

La théorie des questionnaires nous enseigne qu'il existe un seuil minimum que l'on ne peut pas dépasser dans cette économie du codage. La valeur de ce seuil est la quantité d'entropie du fichier :

$$-\sum_{i=1}^{N} f_i \log f_i : f_i \quad \text{fréquence d'appel de l'individu } i.$$

La théorie des questionnaires donne aussi des algorithmes pour constituer les codes optimaux. Voici un exemple numérique emprunté à A. M. Yaglom et I. M. Yaglom, *Probabilité et Information*, Dunod.

Considérons un fichier statistique de base, de 18 individus, dont on connaisse la fréquence (ou une estimation de la fréquence) d'appel de chaque individu : $f_i : i = 1$ à 18. Cherchons un principe de codage optimisé de la zone-identifiant de l'individu. L'exemple numérique est traité en base 2 (codage binaire) afin d'allonger les codes et que l'illustration numérique du gain soit ainsi visible malgré le très faible nombre d'individus du fichier mais le résultat serait vrai, bien entendu, pour toute base de codage.

N° d'ordre de l'individu	Fréquence d'interrogation	Code habituel (longueur fixe)	Code de Shannon-Fano (longueur variable)	Questionnaire du code de Shannon-Fano (équipartition maximum c'est-a-dire information traitée par chaque question maximum)
1	0,3	00000	11	
2	0,2	00001	10	
3	0,1	00010	011	
4	0,1	00011	0101	
5	0,05	00100	0100	
6	0,03	00101	00111	
7	0,03	00110	00110	
8	0,03	00111	00101	
9	0,03	01000	00100	
10	0,03	01001	00011	
11	0,02	01010	000101	
12	0,02	01011	000100	
13	0,01	01100	000011	
14	0,01	01101	0000101	
15	0,01	01110	0000100	
16	0,01	01111	000001	
17	0,01	10000	0000001	
18	0,01	10001	0000000	
	Entropie (en base 2) 3,25	Temps moyen 5	Temps moyen 3,29	

Nota : Le code de longueur variable, par sa structure de questionnaire arborescent, assure un décodage unique. Aucun code n'est une sous-chaîne à gauche d'un autre code.

J. LORIGNY
Administrateur de l'I.N.S.E.E.