Application de la théorie mathématique de l'information pour l'élaboration de questionnaires

Fouad Lazhar Rahmani et Ahmed Chibat

Laboratoire des mathématiques appliquées et modélisation,
Université Frères Mentouri, Constantine 25000, Algérie.

Accepté le 13/12/2009

ملخص

في هذا المقال ندرس وسائل للقياس الكمي المعلوماتي التي قدمتها المراقبة في تحديد القوانين التي تحكم الظواهر العشوائية. من هذه الدراسة على متغير،نبني تصور حصول على معلومات عن مفهوم الكون نسبي. نظهر أنه خلال تنقيح الدراسة التي أجراها حيث تفكك الطابع ، ثمة عتبة الاحتمالات المرتبطة بطرائق مختلفة. هذه العتبة تحدد الحالات التي يكون فيها الحصول على معلومات دائمة وتلك التي كان من قبيل الوهم. نظهر كيف يمكن تمديد هذه الدراسة على حالة العديد من المتغيرات. نحصل عليها من وسائل كمي للاختيار، خطوة خطوة، والمتغيرات، واحترام مبدأ الكون الأقصى. ويمكن لهذا الأسلوب في نتائج التنمية، في مرحلة ما قبل الدراسة، والاستبيانات شحيح جدا الحصاد المعلومات النصيب الأكبر.

الكلمات المفتاحية: *شانون للمعلومات ؛انتروبيا نسبية؛انتروبيا قصوى؛ مجموعات.*

Résumé

Dans cet article nous étudions le moyen de mesurer quantitativement l'information apportée par l'observation lors de l'identification des lois qui régissent les phénomènes aléatoires. A partir de l'étude sur une variable, nous construisons le concept de gain d'information sur la notion d'entropie relative. Nous démontrons que, lors de l'affinement de l'étude par désagrégation des modalités du caractère, il existe un seuil pour les probabilités rattachées aux différentes modalités. Ce seuil détermine les situations où le gain d'information est définitif et celles où il est illusoire. Nous montrons comment cette étude peut s'étendre au cas de plusieurs variables. Nous en déduisons une méthode quantitative de sélection, pas à pas, de variables, respectant le principe de l'entropie maximale. Cette méthode aboutit à l'élaboration, après pré enquête, de questionnaires parcimonieux susceptibles de récolter la plus grande part d'information.

Mots clés : *shannon information; entropie relative; entropie maximum; clustering.*

Abstract

In this paper we study the means of measuring quantitatively the information brought by the observation during the identification of the laws which govern the random phenomena. From the study on a single/one variable, we build the concept of information gain on the relative entropy. We show that, at the time of the refinement of the study by disintegration of the states of the character, there is a threshold for the probabilities attached to the various states. This threshold determines the situations where the gain of information is final and those where it is illusory. We show how this study can be extended to the case of several variables. We deduce from it a quantitative method of selection of variables, step by step, respecting the principle of maximum entropy. This method leads to the development, after pre investigation, of parsimonious questionnaires likely to collect the greatest part of information.

Key words: *shannon information; relative entropy; maximum entropy; clustering.*

Auteur correspondant: achibat@gmail.com (Ahmed Chibat)

1. INTRODUCTION

Tel que relatée par Segal [1], mesurer quantitativement l'information apportée par un échantillon est une préoccupation qui a commencé depuis les travaux de Fisher [2-5]. La définition de l'information s'inscrit au sein d'une structure statistique paramétrée et concerne avant tout ce que l'échantillon peut enseigner à propos de la valeur du paramètre. En 1948 une théorie mathématique de l'information, en tant que notion probabiliste, a vu le jour [6,7]. Cette théorie a suscité beaucoup d'attention et a fait l'objet de plusieurs de recherches [8-10].

Les liens entre les différentes définitions de l'information ont fait l'objet de recherches inaugurant une théorie de l'information unifiée, à la fois statistique et probabiliste Schützenberger [11-16]. Kullback et Leibler [17], dans leur travail de réinterprétation des statistiques autour de la notion d'information, produisent la définition de l'information moyenne apportée par un échantillon en faveur d'une hypothèse H_1 contre une hypothèse H_2 et la définition de la divergence, ou entropie relative, entre deux distributions de probabilités.

Dans notre travail, nous nous plaçons dans le cas où nous ne possédons aucune information préalable et, partant du principe de la raison insuffisante, nous admettons que la première distribution de probabilités est la distribution uniforme. La deuxième distribution est celle qui régi le phénomène aléatoire et que nous cherchons à identifier. De cette manière, l'entropie relative s'identifie, à la limite, avec l'information apportée par l'échantillon. Nous étudions le moyen de concevoir les différentes modalités des variables afin de respecter le principe du maximum de l'entropie [18]. Comme résultat pratique, nous cherchons à aboutir à une méthode de classification descendante à croissance d'information quasi-monotone et à une méthode de conception de questionnaires évitant le plus possible l'acquisition d'information illusoire.

2. ENTROPIE DE LA DISTRIBUTION A PRIORI

Un caractère se répartit en modalités qui peuvent être en nombre fini ou infini. Le caractère peut être une variable aléatoire continue. Dans certains cas, c'est l'observation qui permet d'identifier les différentes modalités du caractère. Dans d'autres cas, ils peuvent déjà être déterminés avant l'observation. Le rôle de l'observation est de noter les occurrences de chaque modalité.

L'incertitude sur l'issue de l'expérience est fonction des probabilités rattachées aux différentes modalités (dont le nombre est k). Elle est mesurée par la quantité:

$$H(X) = -\sum_{i=1}^{k} p_i Log\, p_i \qquad (1)$$

H(x) qui est appelée entropie de l'expérience. La quantité d'information récoltée à la suite de l'expérience est définie comme étant la part d'incertitude éliminée.

2.1 Propriété

L'entropie H(X) est la quantité totale d'information susceptible d'être récoltée. Elle atteint son maximum lorsque toutes les probabilités $p_i \quad i = 1,...,k$ sont égales.

Elle vaut alors :

$$H(X) = Log\, k \qquad (2)$$

2.2 Démonstration

D'une part,

$$H\left(\frac{1}{k},\dots,\frac{1}{k}\right)=-\sum_{i=1}^{k}\frac{1}{k}.Log\frac{1}{k}=-\frac{k}{k}.Log\frac{1}{k}=Logk \quad (3)$$

D'autre part, nous allons utiliser l'inégalité de Jensen qui s'énonce de la manière suivante : Si f est une fonction concave sur l'intervalle $[a,b]$, et $x_1,\dots,x_k$ sont k valeurs arbitraires de l'argument x, alors pour tous nombres positifs $\lambda_1,\dots,\lambda_k$ dont la somme est égale à 1, on a l'inégalité:

$$\sum_{i=1}^{k}\lambda_i.f(x_i)\le f\left(\sum_{i=1}^{k}\lambda_i.x_i\right) \quad (4)$$

Appliquons cette inégalité de Jensen pour

$$x_i=p_i,\qquad \lambda_i=\frac{1}{k}\ ,\qquad 1\le i\le k,$$

$$f(x)=-xLogx \quad (5)$$

Nous obtenons

$$-\sum_{i=1}^{k}\frac{1}{k}.p_i.\mathrm{Log}p_i\le-\left(\sum_{i=1}^{k}\frac{1}{k}.p_i\right)Log\left(\sum_{i=1}^{k}\frac{1}{k}.p_i\right) \quad (6)$$

Et puisque

$$\sum_{i=1}^{k}p_i=1$$

Il en résulte:

$$\frac{1}{k}H(p_1,\dots,p_k)\le-\frac{1}{k}\left(\sum_{i=1}^{k}p_i\right)Log\left(\frac{1}{k}.\sum_{i=1}^{k}p_i\right)=-\frac{1}{k}.Log\frac{1}{k} \quad (7)$$

C'est-à-dire

$$H(p_1,\dots,p_k)\le-Log\frac{1}{k}=\mathrm{Log}\,k \quad (8)$$

Ainsi, quelle que soit sa distribution de probabilités, une variable aléatoire à k modalités ne peut pas produire une quantité d'information supérieure à $Log\,k$. $Log\,k$ peut être vu comme la capacité d'une variable aléatoire à k modalités. Plus le nombre de modalités est grand et plus cette capacité est grande. Augmenter le nombre de modalités prépare à recevoir une plus grande quantité d'information. Etant logarithmique, la croissance de la capacité est rapide au départ mais elle devient insignifiante au-delà d'un certain rang. A la limite, elle converge vers zéro. Si, avant l'observation, nous ne connaissons du caractère que le nombre k de modalités, sans autre information préalable, la distribution des probabilités est uniforme. $Log\,k$ peut ainsi être également vu comme étant l'entropie rattachée à cette distribution a priori.

3. PRODUCTION DE L'EXPERIENCE

Lorsque l'expérience est entreprise, à chaque modalité M_i se trouvera associée une fréquence relative f_i. Pendant que le nombre d'observation augmente, les fréquences relatives évoluent. Elles se stabilisent graduellement et finissent, en vertu de la loi des grands nombres, par converger vers des constantes $p_i\quad i=1,...,k$ qui sont les probabilités.

L'entropie de l'expérience, dans les étapes intermédiaires, se mesure par :

$$H_f(X)=-\sum_{i=1}^{k}f_i Log\ f_i \quad (9)$$

Cette quantité, évoluant suivant l'évolution des $f_i\quad i=1,...,k$ converge vers la quantité:

$$H(X)=-\sum_{i=1}^{k}p_i Log\ p_i \quad (10)$$

4. NOTION DE GAIN D'INFORMATION

L'entropie a posteriori H(X) (eq. 10) est nécessairement inférieure à l'entropie a priori :

$$H(X) \leq Log\, k \tag{11}$$

L'incertitude a priori, qui est l'incertitude totale, ne peut pas être entièrement éliminée par l'expérience. Elle est la somme de deux incertitudes dont une seulement peut disparaître grâce à l'expérience. La partie incompressible est l'entropie H(X) spécifique à la distribution que nous cherchons à identifier. L'autre partie est l'écart entre H(X) et $Log\, k$ et c'est la quantité possible d'information qui peut être apportée par l'expérience. Nous l'appellerons gain de l'information, GI :

$$GI = Log\, k - H(X) \tag{12}$$

5. CONVERGENCE DU GI

Pour un nombre d'observations N donné, le Gain d'Information est :

$$GI_f = Log\, k - H_f(X) \tag{13}$$

Avec l'augmentation du nombre d'observations, la quantité GI_f converge vers la quantité GI.

$$\lim_{N\to\infty} GI_f = \lim_{N\to\infty} \left[Log\, k - H_f(X)\right] = Log\, k - H(X) \tag{14}$$

GI est en fait la distance de Kullback – Leibler, ou entropie relative, entre la distribution uniforme et la distribution à identifier.

Le GI converge également vers une constante quand le nombre de modalités k tend vers l'infini. Ceci peut être perçu comme une version du théorème central limite. Pour fixer les idées, considérons la loi binomiale B(k,p), k étant le nombre de modalités et p la probabilité de succès dans l'épreuve de Bernoulli rattachée à cette loi binomiale.

Pour chaque k et p fixés, la variable binomiale B(k,p) se trouve associée à un GI qui vaut :

$$\mathrm{GI_k} = Log k - H_k(X) = log k + \sum_{i=1}^{k} \mathrm{C_n^i p^i (1-p)^{n-i}}\, log\left[\mathrm{C_n^i p^i (1-p)^{n-i}}\right] \tag{15}$$

Lorsque k augmente, la convergence de la loi binomiale vers la loi normale entraîne la convergence du GI vers une constante qui est le gain d'information associé à la loi normale. D'une manière générale, il est évident que lorsqu'une loi converge vers une autre loi, son GI converge vers le GI de cette loi.

6. DESAGREGATION DE MODALITES

6.1 Augmentation de l'entropie

Théorème 1: Si X est un caractère à k modalités ayant respectivement les probabilités $p_1, p_2, \ldots, p_k$ et si nous construisons un nouveau caractère X' en désagrégeant l'une quelconque des modalités, M_{i_0}, ayant la probabilité p_{i_0}, en deux modalités M_{i_1} et M_{i_2}, avec pour probabilités respectives p_{i_1} et p_{i_2} telles que $p_{i_0} = p_{i_1} + p_{i_2}$ alors:

$$H(X') \geq H(X) \tag{16}$$

Autrement dit, l'entropie augmente toujours en désagrégeant les modalités [19].

6.2 Evolution du GI

Théorème 2: Il existe une condition pour

que le GI augmente en désagrégeant une modalité M_{i_0} (de probabilité p_{i_0}) en deux modalités M_{i_1} et M_{i_2} de probabilités respectives $c.p_{i_0}$ et $(1-c)p_{i_0}$, où c est un réel arbitraire compris entre 0 et 1.

Cette condition est :

$$p_{i_0} \leq \frac{Log(k+1) - Logk}{c.Log\frac{1}{c} + (1-c).Log\frac{1}{1-c}} \qquad (17)$$

Pour la démonstration, nous avons:

$$GI_{k+1} = Log\,(k+1) - H(X') \qquad (18)$$

et

$$GI_k = Log\,k - H(X)$$

Par soustraction:

$$GI_{k+1} - GI_k = \left[Log\,(k+1) - Log\,k\right] - \left[H(X') - H(X)\right]$$

$$= Log\frac{k+1}{k} - p_{i_0}\left[c.Log\frac{1}{c} + (1-c)Log\frac{1}{1-c}\right] \qquad (19)$$

La différence $(GI_{k+1} - GI_k)$ n'est positive que si

$$p_{i_0} \leq \frac{Log\,(k+1) - Log\,k}{c.Log\frac{1}{c} + (1-c).Log\frac{1}{1-c}} = \alpha \qquad (20)$$

6.2.3 Discussion

Lorsqu'il s'agit d'augmenter le nombre de modalités d'un caractère, nous pouvons rencontrer, à chaque désagrégation, deux situations possibles :

a) La probabilité p_{i_0} rattachée à cette modalité est inférieure à α. Dans ce cas nous obtenons une croissance du GI.

b) La probabilité p_{i_0} est supérieure à α, dans quel cas le GI diminuera.

La décomposition dans le second cas apportera une information négative en augmentant l'incertitude globale.

En multipliant les décompositions des modalités entrant dans le second cas, nous ferons augmenter le GI rapidement et d'une manière illusoire avec l'augmentation de k. Il diminuera par la suite lorsque nous devons désagréger les modalités ayant des probabilités entrant dans le premier cas. Pour éviter cette situation et, disposant d'information partielle, il faut choisir comme modalités à désagréger celles dont les probabilités conduisent à maximiser l'incertitude concernant l'information manquante. Le processus doit être conduit de sorte à désagréger d'abord les modalités dont les probabilités p_{i_0} vérifient la condition $p_{i_0} \geq \alpha$.

Lorsqu'à une étape, il n'y a pas de probabilités vérifiant cette condition, alors l'augmentation du GI, suite à la décomposition, est réellement définitive.

Remarque:

La distance de Kullback – Leibler est vue comme mesure d'hétérogénéité. Le gain d'information GI, qui peut être également écrit sous la forme

$$GI = Log\,k.\prod_{i=1}^{k} p_i^{\,p_i} \qquad (21)$$

Le GI peut servir comme mesure d'hétérogénéité de la distribution de probabilités. Cette quantité est maximale lorsque l'une des probabilités est égale à 1, et elle est minimale lorsque toutes les probabilités sont égales. Elle peut servir comme mesure de dispersion lorsqu'il s'agit d'un caractère qualitatif ordinal.

7. CAS DE PLUSIEURS VARIABLES

Théorème 3: La capacité d'un groupe de N variables binaires est égale à N. Plus généralement, la capacité d'un groupe de

N variables, ayant respectivement $m_1, m_2, ..., m_N$ modalités, est égale à: $\sum_{i=1}^{N} Log\, m_i$

Pour la démonstration, il faut souligner d'abord que la description d'une population avec N variables peut se ramener à la description à l'aide d'une seule variable.

Si les N variables sont binaires, la nouvelle variable possédera 2^N modalités, sa capacité est alors:

$$Log\, 2^N = N$$

Si les N variables ont respectivement $m_1, m_2, ..., m_N$ modalités, la nouvelle variable possédera $m_1.m_2.....m_N$ modalités et sa capacité est

$$\log(m_1\, m_2\, ...m_N) = \sum_{i=1}^{N} \log m_i \qquad (22)$$

En effet, du fait qu'il y ait N variables, chaque individu est identifié par un vecteur à N composantes. Si les variables sont binaires, il existe 2^N vecteurs différents. Chacun de ces vecteurs peut être considéré comme une modalité de la nouvelle variable. Si les variables ont des nombres différents de modalités $m_1, m_2, ..., m_N$, il existe $m_1.m_2.....m_N$ vecteurs différents qui seront autant de modalités pour la nouvelle variable.

8. CLASSIFICATION DU GI

Lorsque nous nous disposons à répartir une population en classes, en disposant pour cela d'un certain nombre V de variables ayant chacune un nombre de modalités déterminé, si notre information préalable, résultant d'une pré enquête, est insuffisante, nous pouvons procéder à une sélection graduelle et parcimonieuse des variables en respectant la discussion précédente et le théorème ci-dessus. La première variable sélectionnée est celle dont la distribution des fréquences relatives est la plus uniforme possible. La deuxième variable, ajoutée à la première conduit à une nouvelle distribution de fréquences relatives (qui seront assimilées à des probabilités).

Sur les (V-1) possibilités de choix, sélectionner celle qui produit la distribution de probabilités vérifiant la condition α ou, sinon, en être la plus proche. Le même procédé permettra de sélectionner une à une les variables suivantes. Nous obtenons de cette manière un classement décroissant des V variables en fonction de leur pouvoir discriminant.

- Dans le cas où notre but est la subdivision de la population sous forme de classification descendante : Si nous désirons répartir cette population en M classes, alors nous nous limitons aux L premières variables (dont les nombres de modalités sont respectivement $m_1, m_2, ..., m_L$) et de sorte que $m_1.m_2.....m_L$ soit le proche de M (par excès). Par le fait que l'information acquise est la moins illusoire possible, une partition conçue de la sorte sera la plus stable.
- Dans le cas où nous désirons élaborer un questionnaire : Nous pouvons nous limiter au nombre de variables qui assure le pourcentage attendu de la quantité d'information à récolter. Si, ultérieurement, le nombre des observations devra augmenter, le déséquilibre entre les fréquences relatives ne peut que s'accentuer. C'est ce qui assurera un supplément d'un gain effectif d'information.

Références

[1] J. Segal, *Théorie de l'information : sciences, techniques et société - de la seconde guerre mondiale à l'aube du XXIe siècle*, Thèse de Doctorat, Faculté d'Histoire de l'Université Lyon II, 1998.

[2] R.A.. Fisher, *On the Mathematical*

Foundations of Theoretical Statistis, Philosophical Transactions of the Royal Society, Vol. 222A, 1922, p. 309-368.

[3] R.A. Fisher, *The Theory of Statistical Estimation*, Proceedings of the Cambridge Philosophical Society, Vol. 22, 1925, p. 700-725.

[4] R.A. Fisher, *Probability Likelihood and Quantity of Information in the logic of Uncertain Inference*, Proceedings of the Royal Society A, Vol. 146, 1934, p. 1-8.

[5] R.A.. Fisher, *Statistical Methods and Scientific Inference*, Adelaïde, 1956.

[6] C.E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, 1948, p. 379-423

[7] C.E. Shannon, *Information Theory*, 12th Ed. Encyclopaedia Britannica, William Benton Publishers, Chicago, 1964.

[8] A.I. Khinchin, *Mathematical Foundations of Information Theory*, Dover Pub. Inc. New York, 1953, 1956, 1957.

[9] A. Kolmogorov, *Logical Basis for Information Theory and Probability Theory*, IEEE Trans. on Information Theory, Vol.14, 1968, p. 662-664.

[10] A. Kolmogorov, *Combinatorial foundations of information theory and the calculus of probabilities,* Russian mathematical Surveys, Vol. 38, 1983, p. 29-40.

[11] M.P. Schutzenberger, *Sur les rapports entre la quantité d'information au sens de Fisher et au sens de Wiener*, Comptes rendus de l'Académie des Sciences, Vol. 232, 1951, p. 925-927.

[12] M.P. Schutzenberger, *Contributions aux applications statistiques de la théorie de l'information*, Publications de l'Institut de Statistique de l'Université de Paris, 1953, Vol.3, 117 p.

[13] M.P. Schutzenberger, *On some Measures of Information used in Statistics*, in Information Theory, Papers read at a Symposium held at the Royal Institution, 12-16 Sept. 1955, Ed. C. Cherry, Butterworths Scientific Publications, London, Academic Press, 1956, p. 18-25.

[14] M.P. Schutzenberger, *Théorie de l'information*, in Information et Connaissance, Recherches Interdisciplinaires du Collège de France, sous la Direction de A. Lichnerowicz et F. Perroux, Maloine, Paris, 1983.

[15] S. Kullback, *Note on Information Theory*, Journal of Applied Physics, Vol. 24, 1953, p. 106-107.

[16] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, 2ème édition augmentée, New York, 1968.

[17] S. Kullback, R.A. Leibler, *On Information and Sufficiency*, Annals of Mathematical Statistics, Vol. 22, 1951, p. 79-86.

[18] E.T. Jaynes, *Information Theory and Statistical Mechanics* in The Maximum Entropy Formalism, R.D. Levine et M. Tribus, Eds, Cambridge, M.I.T. Press, 1979.

[19] M. Volle, *Analyse des données*, 3ème édition, Eds. ECONOMICA, Paris, 1993.