Mesure d'entropie asymétrique et consistante

Djamel A. Zighed*, Simon Marcellin*
Gilbert Ritschard**

*Université Lumière Lyon 2, Laboratoire ERIC
{abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
http://eric.univ-lyon2.fr
**Université de Genève, Département d'économétrie, Suisse
gilbert.ritschard@unige.ch

Résumé. Les mesures d'entropie, dont la plus connue est celle de Shannon, ont été proposées dans un contexte de codage et de transmission d'information. Néanmoins, dès le milieu des années soixante, elles ont été utilisées dans d'autres domaines comme l'apprentissage et plus particulièrement pour construire des graphes d'induction et des arbres de décision. L'usage brut de ces mesures n'est cependant pas toujours bien approprié pour engendrer des modèles de prédiction ou d'explication pertinents. Cette faiblesse résulte des propriétés des entropies, en particulier le maximum nécessairement atteint pour la distribution uniforme et l'insensibilité à la taille de l'échantillon. Nous commençons par rappeler ces propriétés classiques. Nous définissons ensuite une nouvelle axiomatique mieux adaptée à nos besoins et proposons une mesure empirique d'entropie plus flexible vérifiant ces axiomes.

1 Introduction

Dans les méthodes qui génèrent des règles de décision du type **Si** *condition* **Alors** *Conclusion* comme les arbres de décision (Breiman et al., 1984; Quinlan, 1993), les graphes d'induction (Zighed et Rakotomalala, 2000),... les mesures d'entropie sont fréquemment utilisées. Or celles-ci reposent sur de nombreuses hypothèses implicites qui ne sont pas toujours justifiées.

Les mesures d'entropie ont été définies mathématiquement par un ensemble d'axiomes en dehors du contexte de l'apprentissage machine. On peut trouver des travaux détaillés dans Rényi (1960), et Aczél et Daróczy (1975). Leur transfert vers l'apprentissage s'est fait de manière peut-être hâtive et mérite d'être revu en détail.

Le présent travail examine et discute des propriétés des entropies dans le cadre des arbres d'induction.

Dans la section suivante, nous fixons quelques notations et rappelons le contexte d'utilisation des mesures d'entropie. Dans la section 3, nous présentons les mesures d'entropie et discutons leurs propriétés et leurs conséquences dans les processus d'induction. Dans la section 4, nous proposons une axiomatique conduisant à une nouvelle mesure d'entropie.

2 Notations, définitions et concepts de base

Nous nous plaçons dans le cadre des arbres de décision qui font explicitement appel aux entropies pour mesurer la qualité de la partition induite en apprentissage.

Soit Ω la population concernée par le problème d'apprentissage. Le profil de tout individu ω de Ω est décrit par p variables, $X_1, \ldots, X_p$, dites variables exogènes ou variables explicatives. Ces variables peuvent être qualitatives ou quantitatives.

Nous considérons également la variable à prédire C, parfois appelée variable endogène ou variable classe ou encore variable réponse. L'ensemble des valeurs prises par cette variable sur la population est un ensemble discret et fini noté $\mathcal{C}$. On note par m_j le nombre de valeurs différentes prises par X_j et par n le nombre de modalités de C. Ainsi, $\mathcal{C} = \{c_1, \ldots, c_n\}$. Et s'il n'y a pas d'ambiguïté, on notera la classe c_i simplement par i.

L'objectif d'un algorithme d'induction d'arbre est de générer un modèle $\phi(X_1, \ldots, X_p)$ de prédiction de C que l'on représente par un arbre de décision. Chaque branche de l'arbre représente une règle. L'ensemble des règles forme le modèle de prédiction qui permet de calculer, pour un nouvel individu dont on ne connaît que les variables exogènes, l'état de la variable endogène. Le développement de l'arbre s'effectue selon un schéma simple : l'ensemble d'apprentissage Ω_a est segmenté itérativement, à chaque fois selon une des variables exogènes $X_j; j = 1, ...p$ de sorte à engendrer la partition de plus faible entropie sur la distribution de C. Les sommets obtenus à chaque itération définissent une partition sur Ω_a. Plus l'arbre grandit, plus la partition devient fine. Le sommet à la racine de l'arbre représente la partition grossière.

Chaque sommet s d'une partition S est caractérisé par une distribution de probabilités des modalités de la variable endogène $C : p(i/s); i = 1, \ldots, n$.

Dans les arbres d'induction, l'entropie H sur la partition S à minimiser est généralement une entropie moyenne calculée comme suit :

$$H(S) = \sum_{s \in S} p(s) h(p(1/s), \ldots, p(i/s), \ldots, p(n/s)) \tag{1}$$

où $h(p(1/s), \ldots, p(i/s), \ldots, p(n/s))$ est par exemple l'entropie de Shannon dont l'expression est donnée plus loin, et $p(s)$ la proportion de cas dans le sommet s.

3 Mesures d'entropie

3.1 Point historique

Le concept d'entropie a été introduit par Hartley (1928) mais fut réellement promu et développé par Shannon (1948) et Shannon et Weaver (1949) dans les années 1940. Ils ont proposé une mesure d'information qui est l'entropie générale d'une distribution finie de probabilités. Suivant le théorème qui caractérise l'entropie de Shannon, plusieurs auteurs comme Khinchin (1957), Forte (1973) et Aczél (1973) ont fondé une axiomatique.

3.2 Entropie de Shannon

Soit une expérience E avec les événements possibles $e_1, e_2, \ldots, e_n$ de probabilités respectives $p_1, p_2, \ldots, p_n$. On suppose que $\sum_i^n p_i = 1$ et $p_i \geq 0$ pour $i = 1, \ldots, n$. L'entropie de

Shannon de la distribution de probabilité est donnée par la formule :

$$H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2}$$

Par continuité on pose $0 \log_2 0 = 0$. D'autres mesures d'entropie existent (Zighed et Rakotomalala, 2000).

3.3 Propriétés théoriques des mesures d'entropie

On considère que $(p_1, p_2, \ldots, p_n)$ pour $n \geq 2$ est pris dans un ensemble fini de distributions de probabilités et on considère le simplexe d'ordre n

$$\Gamma_n = \{(p_1, p_2, \ldots, p_n) : \sum_{i}^{n} p_i = 1; p_i \geq 0\} \tag{3}$$

Une mesure d'entropie est définie comme suit :

$$h : \Gamma_n \rightarrow \mathbb{R} \tag{4}$$

avec les propriétés suivantes :

Non négativité

$$h(p_1, p_2, \ldots, p_n) \geq 0 \tag{5}$$

Symétrie L'entropie est insensible à toute permutation au sein d'un vecteur $(p_1, \ldots, p_n)$ de Γ_n.

$$h(p_1, p_2, \ldots, p_n) = h(p_{\sigma_{(1)}}, p_{\sigma_{(2)}}, \ldots, p_{\sigma_{(n)}}) \tag{6}$$

où σ est une permutation quelconque sur $(p_1, p_2, \ldots, p_n)$.

Minimalité S'il existe k tel que $p_k = 1$ et $p_i = 0$ pour tout $i \neq k$ alors

$$h(p_1, p_2, \ldots, p_n) = 0 \tag{7}$$

Maximalité

$$h(p_1, p_2, \ldots, p_n) \leq h(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \tag{8}$$

Stricte concavité La fonction $h(p_1, p_2, \ldots, p_n)$ est strictement concave.

Ainsi, l'évaluation de l'entropie h d'une partition S nécessite la connaissance de $p(i/s); i = 1, \ldots, n; \forall s \in S$.

4 Mesure d'entropie pour l'apprentissage inductif

Les propriétés des mesures d'entropie que l'on vient de lister ne nous paraissent pas adaptées à l'apprentissage inductif. En effet, d'une part l'incertitude maximale ne correspond pas nécessairement à la distribution uniforme, ainsi dans le cadre de la détection de transactions frauduleuses peu fréquentes il peut par exemple être opportun de conclure à une fraude dès

que la probabilité de celle-ci dépasse un seuil de disons 10%, voire moins. D'autre part, dans la pratique, le calcul de l'entropie repose sur des probabilités estimées et devrait donc tenir compte de leur précision et donc de la taille de l'échantillon. C'est pourquoi nous proposons une nouvelle axiomatique que nous justifions très brièvement et dont l'objectif est d'aboutir à une entropie, que l'on pourrait qualifier d'empirique, qui tient mieux compte de ces considérations pratiques.

4.1 Propriétés requises

Soit $\hbar$ la nouvelle fonction d'entropie que nous voulons bâtir. Nous voulons qu'elle soit empirique, c'est-à-dire fonction des fréquences $f(i/.)$, sensible à la taille N de l'échantillon sur lequel elles sont calculées et qu'elle soit également paramétrée par une distribution $W = (w_1, \ldots, w_j, \ldots, w_p)$ où elle sera maximale.

$$\hbar : \mathbb{N}^* \times \Gamma_n^2 \rightarrow \mathbb{R}^+ \tag{9}$$

On notera, pour une distribution W fixée, $\hbar_W(N, f_1, \ldots, f_i, \ldots, f_n)$.

Nous souhaitons que $\hbar$ possède les propriétés suivantes :

P1 : Non négativité La fonction $\hbar$ doit être à valeur non négative

$$\hbar_W(N, f_1, \ldots, f_j, \ldots, f_n) \geq 0 \tag{10}$$

P2 : Maximalité Soit $W = (w_1, w_2, \ldots, w_n)$ une distribution fixée par l'utilisateur comme étant la moins souhaitée et donc d'entropie maximale. Ainsi, pour N fixé,

$$\hbar_W(N, f_1, \ldots, f_n) \leq \hbar_W(N, w_1, \ldots, w_n) \tag{11}$$

pour toute distribution $(f_1, \ldots, f_n)$ de taille n.

P3 : Asymétrie La nouvelle propriété de maximalité remet en cause l'axiome de symétrie requis par les entropies classiques. Par conséquent, certaines permutations σ pourraient affecter la valeur de l'entropie : $\hbar(f_1, \ldots, f_n) \neq \hbar(f_{\sigma_1}, \ldots, f_{\sigma_n})$. On peut facilement identifier les conditions dans lesquelles la symétrie serait conservée comme par exemple les cas où certains w_i seraient identiques et a fortiori dans le cas de la distribution uniforme.

P4 : Minimalité Dans le contexte classique, l'entropie est nulle dans le cas où la distribution est concentrée en une seule classe, les autres étant vides, c'est-à-dire lorsqu'il existe j tel que $p_j = 1$ et que $p_i = 0$ pour tout $i \neq j$. Cette propriété doit en effet demeurer valide sur le plan théorique. Seulement, en apprentissage ces probabilités sont inconnues. Il serait quand même gênant de dire que l'entropie est nulle dès lors que la distribution est concentrée en une classe. Il faut prendre en considération la taille de l'échantillon qui sert à estimer les p_j. On exige simplement que l'entropie d'une distribution empirique pour laquelle il existe j tel que $f_j = 1$, tende vers 0 quand N devient grand, soit

$$\lim_{N \rightarrow \infty} \hbar_W(N, 0, \ldots, 0, 1, 0 \ldots, 0) = 0 \tag{12}$$

P5 : Consistance Pour un W donné et à distribution fixée, l'entropie devrait être plus faible sur un effectif plus grand.

$$\hbar_W(N+1, f_1, \ldots, f_j, \ldots, f_n) \leq \hbar_W(N, f_1, \ldots, f_j, \ldots, f_n) \tag{13}$$

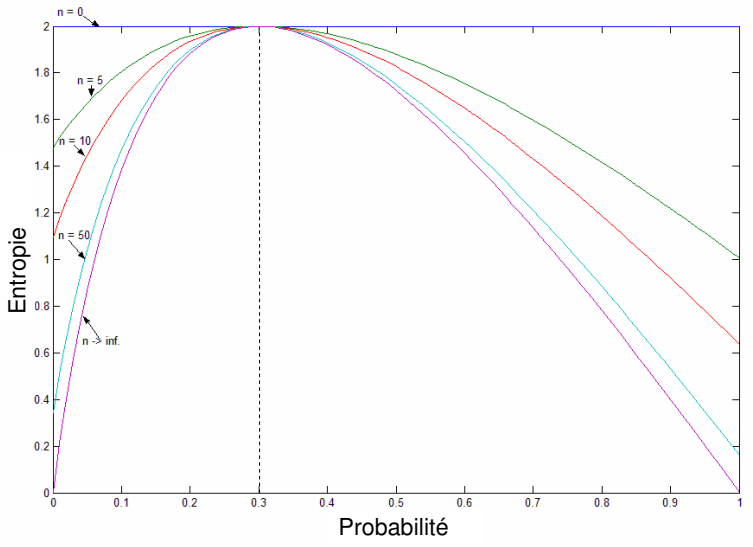

FIG. 1 – *Entropie pour l'apprentissage inductif, 2 classes et* $w = 0.3$

4.2 Formules d'entropie pour apprentissage inductif

On note $\lambda_i = \frac{Nf_i+1}{N+n}$ l'estimateur de Laplace de p_i. Soit $W = (w_1, w_2, \ldots, , w_n)$ le vecteur d'entropie maximale et N la taille de l'échantillon. Nous donnons sans démonstration le théorème suivant :

Théorème

$$h_W(N, f_1, f_2, \ldots, f_n) = \sum_{i=1}^{n} \frac{\lambda_i(1-\lambda_i)}{(-2w_i+1)\lambda_i + w_i^2}$$

est une mesure d'entropie pour processus d'apprentissage inductif qui vérifie les propriétés P1 à P5.

Le graphique 1 visualise la forme de cette entropie pour deux classes avec un vecteur de maximalité $W = (0.3; 0.7)$ et différentes valeurs de $N = 5, 10, 50, ...$

5 Conclusion

Dans ce travail nous avons défini une nouvelle fonction d'entropie qui possède, sur un plan théorique et algorithmique de bonnes propriétés. Cette fonction repose sur le choix du paramètre W. L'idée la plus simple pour fixer W est de prendre la distribution a priori observée sur l'échantillon d'apprentissage. En effet, si l'utilisateur met en œuvre des techniques d'apprentissage c'est pour s'éloigner le plus possible de la distribution a priori. Il est donc naturel qu'elle soit associée à la plus forte entropie. Faute de place, nous n'avons pas pu reporter toutes les critiques que nous pouvons formuler sur l'utilisation des entropies classiques en apprentissage. L'intérêt de notre approche est qu'elle permet de répondre de façon assez simple et sans trop de

"bidouillage" à de nombreux problèmes comme la sous représentation des classes, l'asymétrie des coûts des classes, la sensibilité à la taille de l'échantillon qui évite le sur-apprentissage et l'élagage comme dans CART par exemple.

Références

Aczél, J. (1973). On Shannon inequality, optimal coding, and characterization of Shannon's and Rényi's entropies. In *On Information Theory. 1st Alta Mat. Roma. Symp. Math.*, Volume XVIII, New York. Academic Press.

Aczél, J. et Z. Daróczy (1975). *On measures of information and their characterizations*. New York: Academic Press.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Forte, B. (1973). Why Shannon entropy? In *On Information Theory. 1st Alta Mat. Roma. Symp. Math.*, Volume XVIII, New York. Academic Press.

Hartley, R. V. (1928). Transmission of information. *Bell System Technological Journal* (7), 535–563.

Khinchin, A. I. (1957). *Mathematical foundation of information theory*, Chapter Entropy Concept in Probability Theory, pp. 1–28. New York: Dover Pub. Inc.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Rényi, A. (1960). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1, Berkeley, pp. 547–561. University of California Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal* (27), 379–423, 623–656.

Shannon, C. E. et W. Weaver (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Zighed, D. A. et R. Rakotomalala (2000). *Graphes d'induction : apprentissage et data mining*. Paris : Hermes Science Publications.

Summary

Initially, entropy measures such as Shannon's entropy have been introduced to deal with coding and information transmission. Since the middle of the sixteens, they have, however, been used in many other fields. For instance, in machine learning they are nowadays widely used in decision trees and induction graphs. Nevertheless, entropy measures are not always well suited for building reliable explanatory or prediction models. Their limitations lie mainly in their properties, especially the maximum value that is necessarily reached at the uniform distribution and their insensitiveness to the sample size. We start by recalling these classical properties. We then define a new set of best suited axioms and propose a more flexible empirical entropy measure that fits these axioms.