

# Construction d'une entropie décentrée pour l'apprentissage supervisé

Stéphane Lallich\*, Philippe Lenca\*\*, Benoît Vaillant\*\*\*

\*Université Lyon 2, Laboratoire ERIC

5 avenue Pierre Mendès-France

69676 Bron Cedex, France

stephane.lallich@univ-lyon2.fr

<http://eric.univ-lyon2.fr/~lallich/>

\*\*GET - ENST Bretagne - Département LUSSI

CNRS UMR 2872 TAMCIC

Technopôle de Brest Iroise, CS 83818,

29238 Brest Cedex, France

philippe.lenca@enst-bretagne.fr

<http://perso.enst-bretagne.fr/~lenca/>

\*\*\*UBS - IUT de Vannes - Département STID

Laboratoire VALORIA

8, rue Montaigne,

BP 561, 56017 Vannes, France

benoit.vaillant@univ-ubs.fr

**Résumé.** En apprentissage supervisé, de nombreuses mesures sont fondées sur la notion d'entropie. Une caractéristique majeure des entropies est qu'elles sont maximales lorsque la distribution des modalités de la variable de classe est uniforme, ce qui peut être un inconvénient lorsque cette distribution est très éloignée de l'uniformité. Pour traiter ce cas, nous proposons une entropie décentrée qui prend sa valeur maximale pour une distribution donnée. Cette distribution peut être la distribution *a priori* des classes ou une distribution tenant compte des coûts de mauvaise classification ou plus généralement une distribution fixée par l'utilisateur.

## 1 Motivations

En apprentissage supervisé à partir de variables catégorielles, par exemple en induction par arbres, de nombreux algorithmes d'apprentissage utilisent des mesures d'association prédictive fondées sur l'entropie de Shannon (1948). Considérons une variable de classe  $Y$  à  $q$  modalités et un prédicteur catégoriel  $X$  à  $p$  modalités. La fréquence relative conjointe du couple  $(x_i, y_j)$  est notée  $p_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, q$ . En outre, on désigne par  $h(Y)$  l'entropie de Shannon *a priori* de  $Y$ ,  $h(Y) = -\sum_{j=1}^q p_{.j} \log_2 p_{.j}$ , et par  $h(Y/X)$  l'espérance de l'entropie de  $Y$  conditionnellement à  $X$ ,  $h(Y/X) = E(h(Y/X = x_i))$ . Parmi les mesures usuelles fondées sur l'entropie de Shannon, étudiées notamment par Wehenkel (1996), on citera en particulier :

## Entropie décentrée

- le gain d'entropie (Quinlan (1975)), qui vaut  $h(Y) - h(Y/X)$  ;
- le coefficient  $u$  de Theil (1970), qui est le gain relatif d'entropie de Shannon, à savoir le gain normalisé par l'entropie *a priori* de  $Y$ , valant  $\frac{h(Y) - h(Y/X)}{h(Y)}$  ;
- le gain-ratio de Quinlan (1993) qui rapporte le gain d'entropie dû à  $X$  à l'entropie de  $X$  plutôt qu'à l'entropie *a priori* de  $Y$ , afin de pénaliser les prédicteurs ayant le plus de modalités, ce qui correspond à  $\frac{h(Y) - h(Y/X)}{h(X)}$  ;
- le coefficient de Kvalseth (1987), qui normalise le gain d'entropie par la moyenne des entropies de  $X$  et de  $Y$ , valant  $\frac{2(h(Y) - h(Y/X))}{h(X) + h(Y)}$ .

La particularité de ces coefficients est que l'entropie de Shannon d'une distribution est maximale lorsque la distribution est uniforme. Même si c'est le gain d'entropie par rapport à l'entropie *a priori* de  $Y$  qui figure au numérateur de chacun des coefficients précités, les entropies de  $Y$  et de  $Y/X = x_i$  qui interviennent dans ce gain sont évaluées sur une échelle dont le "zéro" est la distribution uniforme des classes.

Il serait plus logique d'apprécier directement le gain d'entropie à l'aide d'une échelle dont le "zéro" serait la distribution *a priori* des classes. Cette caractéristique des coefficients fondés sur l'entropie est particulièrement contestable lorsque les classes à apprendre sont de fréquences très inégales ou lorsque les coûts de classification sont très inégaux.

Nous proposons dans ce papier une version décentrée de l'entropie qui permet d'apprécier directement à quel point le prédicteur candidat permet d'améliorer la distribution de la variable de classe. Après avoir présenté les travaux de référence par rapport au but poursuivi (section 2), nous exposons de façon détaillée le principe de décentrage de l'entropie de Shannon dans le cas d'une variable booléenne (section 3). Nous généralisons ensuite la méthode proposée au cas d'une variable ayant un nombre quelconque de modalités (section 4) et nous montrons comment étendre la démarche au cas d'une entropie généralisée (section 5) pour ensuite conclure (section 6).

## 2 Etat de l'art

Le principe de construction de cette entropie décentrée a été esquissé dans (Lallich et al., 2005) pour le cas où la variable de classe est booléenne. Dans ce précédent travail, nous proposons une version paramétrée de différentes mesures statistiques de l'intérêt des règles d'association du type  $A \rightarrow B$ , en particulier de l'intensité implication entropique (Gras et al., 2001). Pour construire une mesure statistique, Lerman et al. (1981) proposent de procéder comme suit : on commence par choisir une grandeur d'intérêt, par exemple le nombre de contre-exemples de la règle, ainsi qu'un modèle aléatoire et une hypothèse nulle  $H_0$  qui spécifie la valeur de référence  $\theta$  pour apprécier la confiance de la règle. On détermine ensuite la loi de la grandeur d'intérêt sous  $H_0$ . On construit alors une mesure statistique en centrant et réduisant la grandeur d'intérêt sous  $H_0$  ou une mesure probabiliste en calculant le complément à 1 de la p-valeur associée à la mesure statistique. Dans la mesure où l'intensité d'implication entropique est la moyenne géométrique de l'intensité d'implication et d'un indice d'inclusion reposant sur les entropies des variables booléennes  $B/A$  et  $\overline{A}/\overline{B}$ , le paramétrage de l'intensité d'implication entropique exige le paramétrage de l'indice d'inclusion et par là même le décentrage de l'entropie qui doit prendre sa valeur maximale pour  $p_{b/a} = \theta$  et non plus pour  $p_{b/a} = 0.5$ , tel que fait dans l'indice d'inclusion.

Dans une optique différente, directement reliée à la recherche d'une mesure d'association prédictive, tout particulièrement dans le cas des arbres de décision, Zighed, Ritschard et Marcellin ont proposé une entropie asymétrique et consistante. Cette mesure est asymétrique au sens où l'on peut choisir la distribution pour laquelle elle est maximale ; par consistante il faut entendre qu'elle prend en compte la taille de l'échantillon.

Dans un premier papier (Marcellin et al. (2006)), ces auteurs traitent d'abord le cas d'une variable de classe booléenne, de fréquences  $p$  pour  $Y = 1$  et  $1 - p$  pour  $Y = 0$ . Ils rappellent d'abord les propriétés classiques de l'entropie de Shannon de  $Y$ , notée  $h(p) = -p \log_2(p) - (1 - p) \log_2(1 - p)$ . Celle-ci est une fonction réelle non négative de  $p$ , qui, entre autres propriétés, vérifie notamment :

**1. Invariance par permutation des modalités**

$h(p)$  ne change pas lorsque l'on permute les modalités de  $Y$ .

**2. Maximalité**

La valeur de  $h(p)$  est maximale lorsque la distribution de  $Y$  est uniforme, c'est-à-dire de fréquences égales à  $1/2$  pour chacune des deux modalités de  $Y$ .

**3. Minimalité**

La valeur de  $h(p)$  est minimale lorsque la distribution est certaine, concentrée sur l'une des modalités de  $Y$ , toutes les autres modalités étant de fréquence nulle.

**4. Concavité stricte**

L'entropie  $h(p)$  est une fonction strictement concave.

Marcellin et al. (2006) conservent la propriété 4 (*concavité stricte*) mais modifient la propriété 2 de telle sorte que l'entropie soit maximale pour une distribution laissée au choix de l'utilisateur (maximale pour  $p = \theta$ , où  $\theta$  est choisi par l'utilisateur), ce qui impose de renoncer à la propriété 1 (*invariance par permutation des modalités*). Ils proposent :

$$h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p + \theta^2}$$

On observera que pour  $\theta = 0.5$ , cette entropie asymétrique se confond avec l'entropie quadratique de Gini. Dans un second papier, les mêmes auteurs étendent leur approche au cas d'une variable de classe à  $q$  modalités (Zighed et al. (2007)). En outre, dans la mesure où l'on ne peut qu'estimer la distribution réelle  $(p_j)_{j=1,2,\dots,q}$  par la distribution empirique  $(f_j)_{j=1,2,\dots,q}$ , ils souhaitent que pour une même distribution de fréquences empiriques, la valeur de l'entropie soit d'autant plus faible que  $n$  est grand (propriété 5, nouvelle propriété dite de *consistance*). Ils sont ainsi amenés à modifier la propriété 3 (*Minimalité*) en une propriété 3' (*Minimalité asymptotique*) : l'entropie d'une variable certaine est simplement astreinte à tendre vers 0 lorsque  $n \rightarrow \infty$ . Pour satisfaire ces nouvelles propriétés 3' et 5, ils suggèrent d'estimer les fréquences théoriques  $p_j$  par leur estimateur de Laplace,  $\hat{p}_j = \frac{nf_j+1}{n+q}$ . Ils proposent ainsi une entropie asymétrique consistante définie par :

$$h_\theta(p) = \sum_{j=1}^q \frac{\hat{p}_j(1-\hat{p}_j)}{(1-2\theta_j)\hat{p}_j + \theta_j^2}$$

Une des particularités du principe de décentrage que nous proposons dans ce papier, par rapport à celui proposé par Zighed et al. (2007) est qu'au lieu de donner une seule entropie décentrée, il s'adapte à n'importe quel type d'entropie, que ce soit une entropie de Shannon ou plus généralement une entropie d'ordre bêta de Daroczy (Daroczy (1970)).

### 3 Entropie décentrée pour les variables booléennes

#### 3.1 Principe de construction

On considère une variable de classe  $Y$  qui comporte  $q = 2$  modalités. La distribution de fréquences de  $Y$  pour les valeurs 0 et 1 est notée  $(1 - p, p)$ . Nous voulons définir une entropie décentrée associée à  $(1 - p, p)$ , notée  $\eta_\theta(p)$ , qui soit maximale lorsque  $p = \theta$ , où  $\theta$  est fixé par l'utilisateur, et non pas lorsque  $p = 0.5$  (cas d'une distribution uniforme). Pour définir cette entropie décentrée, suivant la démarche décrite dans Lallich et al. (2005), nous proposons de transformer la distribution  $(1 - p, p)$  en  $(1 - \pi, \pi)$ , où :

$$\pi = \frac{p}{2\theta} \text{ si } 0 \leq p \leq \theta, \quad \pi = \frac{p + 1 - 2\theta}{2(1 - \theta)} \text{ si } \theta \leq p \leq 1$$

En toute rigueur, les fréquences transformées devraient être notées  $1 - \pi_\theta$  et  $\pi_\theta$ . Elles sont notées  $1 - \pi$  et  $\pi$  dans un souci de simplification. Ce sont bien des fréquences, soit  $0 \leq \pi \leq 1$ . L'entropie décentrée  $\eta_\theta(p)$  est alors définie comme l'entropie de  $(1 - \pi, \pi)$  :

$$\eta_\theta(p) = -\pi \log_2 \pi - (1 - \pi) \log_2(1 - \pi)$$

Par rapport à la distribution  $(1 - p, p)$ , il est clair que  $\eta_\theta(p)$  n'est pas une entropie au sens strict du terme. Ses propriétés doivent être analysées en tenant compte du fait que  $\eta_\theta(p)$  est l'entropie de la distribution transformée  $(1 - \pi, \pi)$ , soit  $\eta_\theta(p) = h(\pi)$ . Le comportement de cette entropie est illustré par la figure 1 pour  $\theta = 0.2$ .

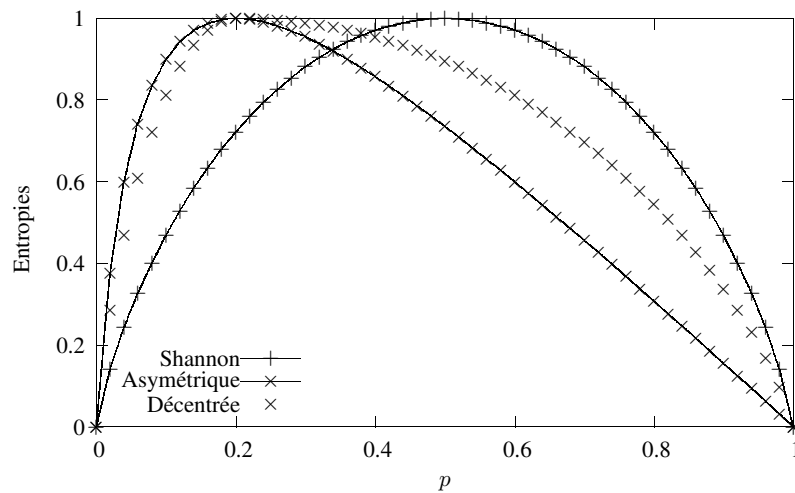


FIG. 1 – Entropies décentrée, asymétrique et de Shannon

### 3.2 Propriétés

L'entropie décentrée conserve différentes propriétés de l'entropie, parmi celles étudiées notamment par Zighed et Rakotomalala (1998) dans un contexte de Data mining. Ces propriétés sont faciles à démontrer dans la mesure où  $\eta_\theta(p)$  est définie comme une entropie en fonction de  $\pi$  et en possède les caractéristiques.

Au préalable, pour démontrer certaines des propriétés de  $\eta_\theta(p)$  en fonction de  $p$ , il faut calculer ses dérivées première et seconde par rapport à  $p$  sachant que  $\eta_\theta(p) = h(\pi)$  est une fonction concave de  $\pi$  (entropie) où  $\pi$  est une fonction linéaire croissante par morceaux de  $p$ . Pour faciliter ces calculs, on considère la fonction  $\eta(x) = h(f(x))$ , où  $h$  est concave et  $f$  est linéaire croissante,  $f(x) = ax + b$ ,  $a > 0$ , soit  $f'(x) = a$  et  $f''(x) = 0$ . Alors les dérivées première et seconde de  $\eta(x)$ , par rapport à  $x$ , s'écrivent respectivement  $\eta'(x) = h'(f(x))f'(x) = ah'(f(x))$  et  $\eta''(x) = a^2h''(f(x))$ .

#### 1. Invariance par permutation des catégories

Cette propriété des entropies est volontairement abandonnée, puisque l'on décentre l'entropie.

#### 2. Maximalité

$\eta_\theta(p)$  est maximale, de valeur 1, pour  $\pi = 0.5$ , soit  $p = 0.5 \times 2\theta = \theta$ . Sa dérivée première par rapport à  $\theta$  s'écrit :

$$\begin{aligned} - \eta'_\theta(p) &= \frac{1}{2\theta} h'(\pi) = \frac{1}{2\theta} (\log_2(1 - \pi) - \log_2 \pi), \text{ pour } 0 \leq p \leq \theta, \\ - \eta'_\theta(p) &= \frac{1}{2(1-\theta)} h'(\pi) = \frac{1}{2(1-\theta)} (\log_2(1 - \pi) - \log_2 \pi), \text{ pour } \theta \leq p \leq 1 \end{aligned}$$

La dérivée est nulle pour  $\pi = 0.5$ , soit  $p = \theta$ . L'entropie décentrée  $\eta_\theta(Y)$  est donc une fonction dérivable en tout point qui prend la valeur maximale 1 pour  $p = \theta$

#### 3. Minimalité

$\eta_\theta(p)$  est minimale pour  $\pi = 0$  et  $\pi = 1$ , donc pour  $p = 0$  et  $p = 1$ .

#### 4. Concavité

D'après le calcul préalable :

$$\begin{aligned} - \eta''_\theta(p) &= \frac{1}{4\theta^2} h''(\pi) = \frac{-1}{4\theta^2 L n 2} \frac{1}{\pi(1-\pi)}, \text{ pour } 0 \leq p \leq \theta, \\ - \eta''_\theta(p) &= \frac{1}{4(1-\theta)^2} h''(\pi) = \frac{-1}{4(1-\theta)^2 L n 2} \frac{1}{\pi(1-\pi)}, \text{ pour } \theta \leq p \leq 1 \end{aligned}$$

Par suite,  $\eta_\theta(p)$  est une fonction concave de  $p$ . On remarquera qu'au point  $p = \theta$ , la dérivée seconde à gauche est distincte de la dérivée seconde à droite.

### 3.3 Autres propriétés

#### 1. Prise en compte de la taille de l'échantillon (consistance) et minimalité asymptotique.

Pour satisfaire cette propriété introduite par Zighed et Rakotomalala (1998) pour les arbres de décision, deux possibilités s'offrent à nous, que nous n'avons pas encore exploitées. En premier lieu, on peut suivre la démarche utilisée par Zighed et al. (2007) pour construire une entropie asymétrique consistante, qui consiste à estimer les fréquences théoriques à l'aide de l'estimateur de Laplace. Par ailleurs, on peut avoir recours à la méthode delta (Goodman et Kruskal (1972)) pour estimer la variance asymptotique de l'entropie et raisonner sur la borne basse de l'intervalle de confiance.

#### 2. Condition de transfert (Pigou-Dalton).

| $p$ | $h(p)$ | $\eta_\theta(p)$ |
|-----|--------|------------------|
| 0   | 0      | 0                |
| 0.1 | 0.469  | 0.811            |
| 0.2 | 0.722  | 1                |
| 0.3 | 0,881  | 0.989            |
| 0.4 | 0.971  | 0.954            |
| 0.5 | 1      | 0.896            |
| 0.6 | 0.971  | 0.811            |

TAB. 1 – Quelques valeurs remarquables de  $h(p)$  et  $\eta_\theta(p)$ .

On sait que l'entropie de Shannon vérifie la condition de transfert (dite de Pigou-Dalton) au sens où elle augmente lorsqu'une modalité plus fréquente transfère une partie de sa masse de fréquence sur l'autre modalité sans que l'ordre soit modifié. Pour l'entropie décentrée, comme l'illustre la figure 1, cette condition de transfert reste vraie pourvu que la position de chacune des deux modalités par rapport à  $\theta$  ne soit pas modifiée. Le tableau 1 illustre ce phénomène (dans le cas  $\theta = 0.2$ ). Si l'on transfère une fréquence de 0.1, sur la modalité de plus faible fréquence, l'entropie  $h(p)$  augmente tant que l'ordre des 2 modalités reste inchangé (chaque modalité doit rester du même côté de 0.5 à l'issue du transfert). Dans le cas de  $\eta_\theta(p)$ , il faut que le transfert conserve la position de chaque modalité par rapport à  $\theta$ .

### 3. Propriétés perdues

Certaines propriétés des entropies semblent perdues, ou sans objet, ainsi l'insensibilité aux modalités de fréquence nulle, l'uniformité partagée, le comportement en cas de scission-fusion de modalités et la pseudo-additivité. Il faut donc souligner que les entropies décentrées ne sont pas des entropies au sens stricte du terme.

## 4 Entropie décentrée pour une variable à $q$ modalités

Pour étendre la définition de l'entropie décentrée au cas d'une variable  $Y$  ayant  $q$  modalités,  $q > 2$ , on utilise une stratégie similaire à celle utilisée dans le cas booléen. On note  $\underline{p} = (p_1, p_2, \dots, p_q)$  le vecteur des fréquences de  $Y$  et  $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_q)$ , le vecteur des fréquences de la distribution de référence, par exemple la distribution *a priori* de  $Y$  en apprentissage supervisé.

### 4.1 Recherche de la forme décentrée

L'entropie de  $\underline{p}$  s'écrit  $h(\underline{p}) = -\sum_{j=1}^q p_j \log_2 p_j$ , alors que l'on recherche l'entropie décentrée sous la forme  $\eta(\underline{p}) = -\sum_{j=1}^q \pi_j \log_2 \pi_j$ , où pour être analogues à des fréquences relatives, les  $\pi_j$  doivent vérifier :

- $0 \leq \pi_j \leq 1$
- $\sum_{j=1}^q \pi_j = 1$

C'est ainsi que :

- $\pi_j$  passe de 0 à  $1/q$ , lorsque  $p_j$  passe de 0 à  $\theta_j$
- $\pi_j$  passe de  $1/q$  à 1, lorsque  $p_j$  passe de  $\theta_j$  à 1

Cherchons les  $\pi_j$  sous la forme  $\pi_j = \frac{p_j - b_j}{a_j}$ .

Dans le cas  $0 \leq p_j \leq \theta_j$  on a :

- $p_j = 0, \pi_j = 0$ , soit  $0 = -\frac{b_j}{a_j}$  et  $b_j = 0$
- $p_j = \theta_j, \pi_j = 1/q$ , soit  $1/q = \frac{\theta_j - b_j}{a_j}$  et  $a_j = q\theta_j$

Il vient :

$$\pi_j = \frac{p_j}{q\theta_j}, \text{ si } 0 \leq p_j \leq \theta_j$$

Dans le cas  $\theta_j \leq p_j \leq 1$  :

- $p_j = \theta_j, \pi_j = 1/q$ , soit  $1/q = \frac{\theta_j - b_j}{a_j}$  et  $a_j = q(\theta_j - b_j)$
- $p_j = 1, \pi_j = 1$ , soit  $1 = \frac{1 - b_j}{a_j}$  et  $a_j = 1 - b_j$

Par suite,  $1 - b_j = q(\theta_j - b_j)$ , d'où  $b_j = \frac{q\theta_j - 1}{q-1}$  et  $a_j = \frac{q(1 - \theta_j)}{q-1}$

On obtient alors :

$$\pi_j = \frac{p_j - b_j}{a_j} = \frac{q(p_j - \theta_j) + 1 - p_j}{q(1 - \theta_j)}$$

A titre de vérification :

- si  $q = 2$ , on retrouve bien les formules qui précèdent (section 3),
- par construction,  $\pi_j$  prend les valeurs 0,  $1/q$  et 1, lorsque  $p_j$  prend les valeurs 0,  $\theta_j$  et 1,
- on a bien  $0 \leq \pi_j \leq 1, j = 1, 2, \dots, q$ ,
- seul problème, la condition de normalisation, qui n'est automatiquement vérifiée que pour  $q = 2$ .

Pour régler ce problème, il suffit de normaliser les  $\pi_j$ . On obtient ainsi les  $\pi_j^*$  définis par  $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^q \pi_j}$ . Les propriétés précitées sont conservées, puisque le facteur de normalisation vaut 1 pour  $p_j = \theta_j$ , car alors les  $\pi_j$  sont égaux à  $1/q$ . L'entropie décentrée pour une variable à  $q$  modalités est alors définie par  $\eta_{\theta}(p) = h(\underline{\pi}^*)$ .

## 4.2 Représentation graphique

Pour illustrer le comportement de l'entropie décentrée, il est possible de représenter les valeurs de  $\eta_{\theta}(p)$  dans le cas de  $q = 3$  catégories. Les fréquences étant liées par la condition de normalisation, on peut représenter l'entropie décentrée par une surface située dans un hyperplan de  $R^4$ .

## 5 Décentrage des entropies généralisées

L'entropie de Shannon n'est pas la seule fonction de diversité ou d'incertitude utilisable pour construire des coefficients d'association prédictive. Déjà, Goodman et Kruskal (1954) avaient proposé une présentation unifiée des trois coefficients usuels que sont le  $\lambda$  de Guttman, le  $u$  de Theil et le  $\tau$  de Goodman et Kruskal, sous l'appellation de coefficient *PRE* (*Proportional Reduction in Error*). De façon plus générale (cf. Lallich (2002) où l'on retrouvera le détail des coefficients cités ici), nous avons construit les coefficients du type *Proportional Reduction in Diversity* (*PRD*), qui sont l'analogie du gain normalisé lorsque l'entropie de Shannon est

## Entropie décentrée

remplacée par une fonction d'incertitude quelconque. Pour qu'une telle construction soit justifiée, comme le note C. d'Aubigny (d'Aubigny, 1980), il suffit que la fonction d'incertitude soit concave, afin que la réduction moyenne de diversité de  $Y$  due au conditionnement suivant  $X$  soit positive, grâce à l'inégalité de Jensen. Si la fonction  $I$  choisie est l'entropie quadratique de Gini,  $I(Y) = 2(1 - \sum_{j=1}^q p_j^2)$  (indice de diversité de Gini-Simpson) le gain relatif correspond au coefficient  $\tau$  de Goodman et Kruskal, alors que si la fonction choisie est  $I(Y) = q - 1$  (indice de diversité du nombre d'espèces, en écologie) le gain relatif correspond au coefficient  $\lambda$  de Guttman, Goodman et Kruskal.

Plus généralement encore, nous avons remarqué que les fonctions d'incertitude utilisables étaient soit des entropies généralisées d'ordre  $\beta$  de Daroczy (1970), soit des diversités de rangs d'ordre  $\rho$  introduites par Patil et Taillie (1982). Nous avons ainsi proposé (Lallich (2002)) une écriture unique pour la quasi-totalité des coefficients usuels sous la forme d'une réduction normalisée d'entropie généralisée ou de diversité de rangs :

$$\lambda_\alpha(Y/X) = \frac{I(Y) - I(Y/X)}{\alpha I(Y) + (1 - \alpha)I(X)}$$

Dans cette formule,  $I$  renvoie aussi bien aux entropies d'ordre  $\beta$  qu'à leur équivalent en termes de diversité de rangs d'ordre  $\rho$ , alors que  $\alpha$  est à la disposition de l'utilisateur pour arbitrer entre les deux normalisations usuelles. Cette expression permet de retrouver les coefficients usuels ( $\alpha = 1$ ) fondés sur une entropie généralisée ( $\beta = 0$  : nombre de catégories ;  $\beta = 1$  : Theil ;  $\beta = 2$  : Gini) ou de rangs ( $\rho = 0$  : Guttman ;  $\rho = 1$  : Utton) ainsi que des analogues du gain-ratio ( $\alpha = 0$ ) et du coefficient de Kvalseth ( $\alpha = 0.5$ ) et d'en générer de nouveaux. La stratégie de décentrage que nous avons proposée s'applique sans difficultés au cas où la fonction d'incertitude choisie est une entropie généralisée ou une entropie de rangs.

Par exemple, la formule générale des entropies généralisées d'ordre  $\beta$  s'écrit  $H_\beta(\underline{p}) = \frac{2^{\beta-1}}{2^{\beta-1}-1} \left(1 - \sum_{j=1}^q p_j^\beta\right)$ . Pour décentrer cette entropie, il faut d'abord transformer les fréquences  $p_j$  en  $\pi_j$ , puis normaliser les  $\pi_j$ , pour obtenir les pseudos-fréquences  $\pi_j^*$ , suivant la procédure décrite dans la section qui précède. On obtient l'évaluation de la distribution  $\underline{p}$  par l'entropie décentrée d'ordre  $\beta$  en formant :

$$\eta_\beta(\underline{p}) = H_\beta(\underline{\pi}^*) = \frac{2^{\beta-1}}{2^{\beta-1}-1} \left(1 - \sum_{j=1}^q \pi_j^{*\beta}\right)$$

La version décentrée des entropies de rangs est construite suivant le même procédé. Par exemple, dans le cas  $\rho = 0$ , qui correspond à la logique du coefficient de Guttman, on a  $H_{\rho=0}(\underline{p}) = 2(1 - \max\{p_j, j = 1, 2, \dots, q\})$ , d'où :

$$\eta_{\rho=0}(\underline{p}) = H_{\rho=0}(\underline{\pi}^*) = 2(1 - \max\{\pi_j^*, j = 1, 2, \dots, q\})$$

Pour illustrer le comportement des entropies généralisées décentrées, nous avons représenté  $\eta_\beta$ , pour  $\beta = 0, 0.5, 1, 2, 5$ , ainsi que  $\eta_{\rho=0}$  (figure 2) et l'entropie asymétrique de Zighed et al. (Zighed et al. (2007)) dans le cas où la distribution *a priori* de la variable de classe est (0.8, 0.2), ce qui correspond à  $\theta = 0.2$ . Les différences de comportement apparaissent clairement sur cette figure qui montre bien l'intérêt et la spécificité de l'entropie asymétrique que nous proposons. C'est en fait un "kit de décentrage" que l'on peut appliquer à n'importe quelle mesure d'association prédictive reposant sur un gain d'incertitude. Le choix de la valeur de  $\beta$  ou  $\rho$  dépend de la réactivité que l'on attend de la mesure.

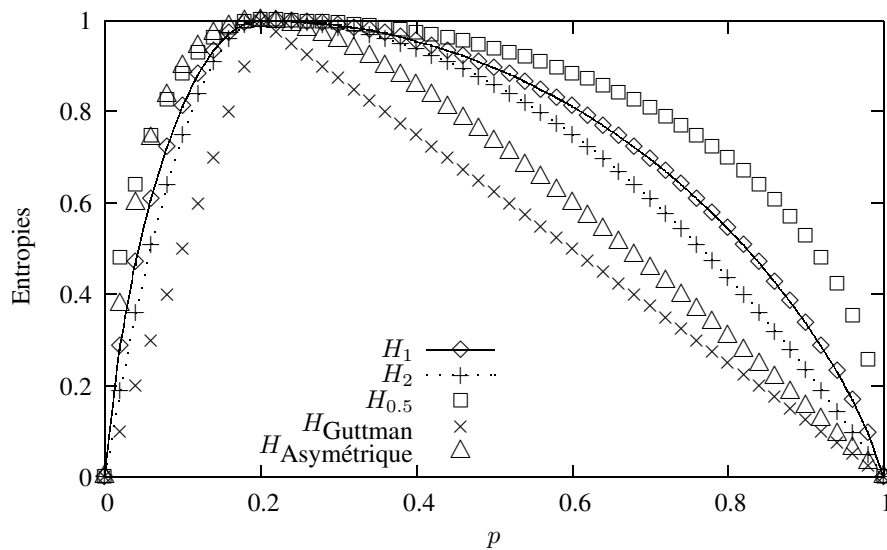


FIG. 2 – Décentrage des entropies généralisées

## 6 Conclusion et travaux futurs

Les mesures d'association prédictive usuelles peuvent être exprimées sous forme d'un gain normalisé associé à une fonction d'incertitude, entropie généralisée ou diversité de rangs. Au terme de cette première étape, nous proposons une méthode de décentrage qui permet d'associer une entropie décentrée à n'importe quelle entropie généralisée d'ordre  $\beta$ , ou entropie de rangs d'ordre  $\rho$ .

La phase suivante de ce travail est bien sûr la mise en œuvre des entropies décentrées sur des données réelles, notamment lorsque celles-ci présentent une variable de classe très déséquilibrée, pour examiner dans quelle mesure elles permettent d'améliorer les performances des algorithmes de classification supervisée.

## Références

- Daroczy, A. (1970). Generalized information functions. *Information and Control* (16), 36–51.
- d'Aubigny, C. (1980). *Etude de la morphologie des indices d'association*. Ph. D. thesis, Thèse de 3e cycle, Université Joseph Fourier, Grenoble, France.
- Goodman, L. A. et W. H. Kruskal (1954). Measures of association for cross classifications, i. *JASA I*(49), 732–764.
- Goodman, L. A. et W. H. Kruskal (1972). Measures of association for cross classifications, iv. *JASA IV*(67), 415–421.

- Gras, R., P. Kuntz, R. Couturier, et F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Conférence Extraction des connaissances et apprentissage (EGC 2001) 1(1-2)*, 69–80.
- Kvalseth, T. O. (1987). Entropy and correlation : some comments. *IEEE Trans. on Systems, Man and Cybernetics 17(3)*, 517–519.
- Lallich, S. (2002). *Mesure et validation en extraction des connaissances à partir des données*. Ph. D. thesis, Habilitation à Diriger des Recherches, Université Lyon 2, France.
- Lallich, S., B. Vaillant, et P. Lenca (2005). Parametrised measures for the evaluation of association rule interestingness. In *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, pp. 220–229.
- Lerman, I., R. Gras, et H. Rostam (1981). Elaboration et évaluation d'un indice d'implication pour données binaires. *Mathématiques et Sciences Humaines 74*, 5–35.
- Marcellin, S., D. Zighed, et G. Ritschard (2006). An asymmetric entropy measure for decision trees. In *Proceedings in Computational Statistics, Berlin : Springer, CD*, pp. 975–982.
- Patil, G. et C. Taillie (1982). Diversity as a concept and its measurement. *Journal of American Statistical Association 77(379)*, 548–567.
- Quinlan, J. (1975). *Machine Learning*, Volume 1.
- Quinlan, J. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal (27)*, 379–423, 623–656.
- Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology (76)*, 103–154.
- Wehenkel, L. (1996). On uncertainty measures used for decision tree induction. In *Proceedings of Info. Proc. and Manag. Of Uncertainty*, pp. 413–418.
- Zighed, D., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. *Actes 7e Conférence EGC, Extraction et Gestion des Connaissances, Namur to appear*.
- Zighed, D. A. et R. Rakotomalala (1998). *Graphes d'induction et apprentissage machine*. Hermès Paris.

## Summary

In supervised learning, many measures are based on the concept of entropy. A major characteristic of the entropies is that they take their maximal value when the distribution of the modalities of the class variable is uniform. To deal with the case where the *a priori* frequencies of the class variable modalities are very unbalanced, we propose a decentered entropy which takes its maximum value for a distribution fixed by the user. This distribution can be the *a priori* distribution of the class variable modalities or a distribution taking into account the costs of misclassification.