Construction of an off-centered entropy for supervised learning

Stéphane Lallich, Philippe Lenca, Benoît Vaillant

Construction of an off-centered entropy for supervised learning

Stéphane Lallich[1], Philippe Lenca[2], and Benoît Vaillant[3]

[1] Université Lyon, Laboratoire ERIC, Lyon 2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France
stephane.lallich@univ-lyon2.fr
[2] GET - ENST Bretagne, CNRS UMR 2872 TAMCIC
Technopôle de Brest Iroise, CS 83818, 29238 Brest Cedex, France
philippe.lenca@enst-bretagne.fr
[3] UBS - IUT de Vannes - Département STID, Laboratoire VALORIA
8, rue Montaigne, BP 561, 56017 Vannes, France
benoit.vaillant@univ-ubs.fr

Abstract. 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 imbalanced, we propose an off-centered 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.
Keywords: supervised learning, entropy, imbalanced class.

1 Motivations

In supervised learning on categorical variables, for example in induction tree, many learning algorithms use predictive association measures based on the entropy proposed by [Shannon, 1948]. Let us consider a class variable Y having q modalities and a categorial predictor X having k modalities. The joint relative frequency of the couple (x_i, y_j) is denoted $p_{ij}, i = 1, 2, \ldots k; j = 1, 2, \ldots q$. What is more, we denote by $h(Y) = -\sum_{j=1}^{q} p_{.j} \log_2 p_{.j}$ the *a priori* Shannon's entropy of Y and by $h(Y/X) = E(h(Y/X = x_i))$ the conditional expectation of the entropy of Y with respect to X. Amongst the usual measures based on Shannon's entropy, studied in particular by [Wehenkel, 1996], we especially wish to point out:

- the entropic gain ([Quinlan, 1975]), which values $h(Y) - h(Y/X)$;
- the u coefficient of [Theil, 1970], which is the relative gain of Shannon's entropy. In other words, it is the entropic gain, normalised on the *a priori* entropy of Y, and thus values $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- the gain-ratio ([Quinlan, 1993]) which relates the entropic gain of X to the entropy of X, rather than to the *a priori* entropy of Y in order to discard the predictors having many modalities. It values $\frac{h(Y)-h(Y/X)}{h(X)}$;

- the [Kvalseth, 1987] coefficient, which normalises the entropic gain by the mean of the entropies of X and Y. It then values $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$.

The peculiarity of these coefficients is that the Shannon's entropy of a distribution reaches its maximum when this distribution is uniform. Even though it is the entropic gain with respect to the *a priori* entropy of Y which is used in the numerator part of the previously mentioned coefficients, the entropies of Y and $Y/X = x_i$ used in this gain are evaluated on a scale for which the "zero" corresponds to the uniform distribution of classes.

It would seem more logical to evaluate directly the entropic gain through the use of a scale for which the "zero" would correspond to the *a priori* distribution of classes. The above-mentioned characteristic of the coefficients based on the entropy is particularly questionable when the classes to be learned have highly imbalanced frequencies in the data, or when the classification costs differ largely.

We propose in this paper an off-centered version of the entropy, which enables one to directly estimate the degree at which a candidate predictor enhances the distribution of the class variable. After having presented the reference works dealing with the goal followed (section 2), we expose in detail the principles of the off-centered Shannon's entropy when used on a boolean variable (section 3). We then generalise the proposed method to the case of a variable having any number of modalities (section 4) and show how to extend the approach to the construction of a generalised off-centered entropy (section 5). We finally conclude (section 6).

2 State of the art

The construction of an off-centred entropy principle is sketched out in the case of a boolean class variable in [Lallich *et al.*, 2005]. In this previous work, we proposed a parameterised version of several statistical measures assessing the interest of association rules of the form $A \rightarrow B$, in particular the inclusion index and the entropic intensity of implication ([Gras *et al.*, 2001]). In order to build a statistical measure which compares the confidence of the rule with a parameter θ instead of the *a priori* probability of B, we constructed an off-centered entropy centered on θ (see section 3).

With an alternative goal, directly related to the construction of a predictive association measure, especially in the context of decision trees, Zighed, Ritschard and Marcellin proposed a consistent and asymmetric entropy. This measure is asymmetric in the sense that one may choose the distribution for which it will reach its maximum ; and consistent since it takes into account the size of the sampling scheme.

In [Marcellin *et al.*, 2006], the authors deal with a boolean class variable, of frequency p for $Y = 1$ and $1 - p$ for $Y = 0$. They recall the classical properties defined on Shannon's entropy, here valuing $h(p) = -p\log_2(p) - (1-p)\log_2(1-p)$. It is a real positive function of p, verifying notably:

1. **Invariance by permutation of modalities**: $h(p)$ does not change when the modalities of Y are permuted.
2. **Maximality**: the value of $h(p)$ reaches its maximum when the distribution of Y is uniform, in other words when each modality of Y has a frequency of $1/2$.
3. **Minimality**: the value of $h(p)$ reaches its minimum when the distribution of Y is sure, centered on one modality of Y, the others being of null frequency.
4. **Strict concavity**: the entropy $h(p)$ is a strictly concave function.

[Marcellin *et al.*, 2006] preserve the *strict concavity* property but alter the *maximality* one in order to let the entropy reach its maximal value for a user chosen distribution (*i.e.* maximal for $p = \theta$, where θ is fixed by the user). This implies revoking the *invariance by permutation of modalities*. They propose:

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

It can be noticed that for $\theta = 0.5$, this asymmetric entropy corresponds to the quadratic entropy of Gini. In [Zighed *et al.*, 2007], the same authors extend their approach to the situation where the class variable has q modalities. What is more, since one may only make an estimation of the real distribution $(p_j)_{j=1,2...,q}$ with an empirical distribution $(f_j)_{j=1,2...,q}$, they wish that for same values of the empirical distribution, the value of the entropy should decrease as n rises (property 5, a new property called *consistency*). They thus are led to modify the third property (*minimality*) in a new property $3'$ (*asymptotic minimality*): the entropy of a sure variable is only required to tend towards 0 as $n \to \infty$. In order to comply with these new properties, they suggest to estimate the theoretical frequencies p_j by their Laplace estimator, $\widehat{p_j} = \frac{nf_j+1}{n+q}$. They thus propose a consistant asymmetric entropy as:

$$h_\theta(p) = \sum\nolimits_{j=1}^{q} \frac{\widehat{p_j}(1-\widehat{p_j})}{(1-2\theta_j)\widehat{p_j} + \theta_j^2}$$

One of the particularities of the off-centering we here propose, compared to the approach proposed by [Zighed *et al.*, 2007] is that rather than defining a single off-centered entropy, it adapts to whichever kind of entropy, may it be Shannon's entropy or more generally to a Daroczy entropy of order beta ([Daroczy, 1970]).

3 Off-centered entropy for boolean variables

3.1 Construction principle

Let us consider a class variable Y made of $q = 2$ modalities. The frequencies distribution of Y for the values 0 and 1 is noted $(1-p, p)$. We wish to define an

off-centered entropy associated with $(1-p, p)$, noted $\eta_\theta(p)$, which is maximal when $p = \theta$, θ being fixed by the user and not necessarily equal to 0.5 (in the case of a uniform distribution). In order to define the off-centered entropy, following the proposition described in [Lallich *et al.*, 2005], we propose that the $(1-p, p)$ distribution should be transformed into a $(1-\pi, \pi)$, distribution, such that:

- π increases from 0 to 1/2, when p increases from 0 to θ;
- π increases from 1/2 to 1, when p increases from θ to 1.

By looking for an expression of π as $\pi = \frac{p-b}{a}$, on both intervals $0 \leq p \leq \theta$ and $\theta \leq p \leq 1$, we obtain:

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

To be precise, the thus transformed frequencies should be denoted as $1 - \pi_\theta$ et π_θ. We will simply use $1 - \pi$ and π for clarity reasons. They do correspond to frequencies, since $0 \leq \pi \leq 1$. The off-centered entropy $\eta_\theta(p)$ is then defined as the entropy of $(1 - \pi, \pi)$:

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

With respect to the distribution $(1-p, p)$, clearly $\eta_\theta(p)$ is not an entropy strictly speaking. Its properties must be studied considering the fact that $\eta_\theta(p)$ is the entropy of the transformed distribution $(1 - \pi, \pi)$, *i.e.* $\eta_\theta(p) = h(\pi)$. The behavior of this entropy is illustrated in figure 1 for $\theta = 0.2$.

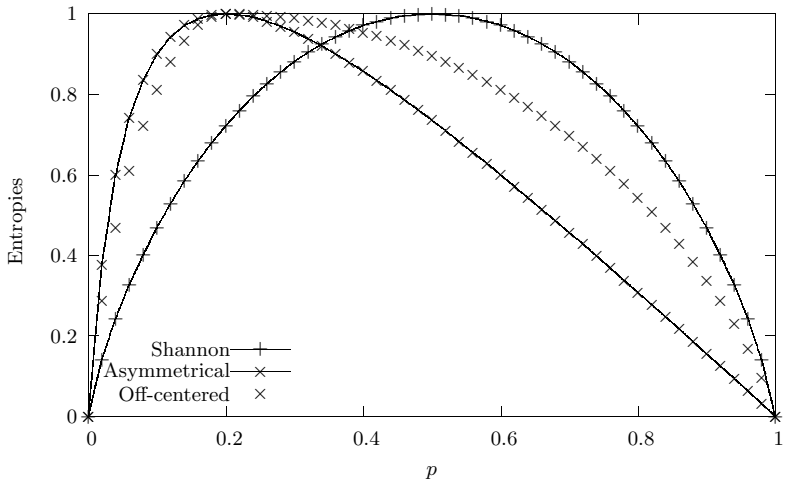

Fig. 1. Off-centered, asymmetric and Shannon's entropies

3.2 Properties

The off-centered entropy preserves various properties of the entropy, among those studied in particular by [Zighed and Rakotomalala, 1998] in a datamining context. Those properties are easy to prove since $\eta_\theta(p)$ is defined as an entropy on π and thus possess such characteristics.

First, in order to prove some of the properties of $\eta_\theta(p)$ in function of p, we must express its first and second derivatives with respect to p, knowing that $\eta_\theta(p) = h(\pi)$ is a concave function of π (entropy), where π is a piecewise linear increasing fonction of p. We thus consider the function $\eta(x) = h(f(x))$, where h is concave, $f(x) = ax + b$ is linear and increasing ($a > 0$, $f'(x) = a$, $f''(x) = 0$). Then, the first and second derivatives of $\eta(x)$ with respect to x are $\eta'(x) = h'(f(x))f'(x) = ah'(f(x))$ and $\eta''(x) = a^2h''(f(x))$.

1. **Invariance by permutation of modalities**: this property is voluntarily abandoned, since we want to construct an off-centered entropy.
2. **Maximality**: $\eta_\theta(p)$ is maximal and values 1 for $\pi = 0.5$, thus when $p = 0.5 \times 2\theta = \theta$. Its first derivative with respect to θ is:
   - $\eta'_\theta(p) = \frac{1}{2\theta}h'(\pi) = \frac{1}{2\theta}\left(\log_2(1-\pi) - \log_2 \pi\right)$, for $0 \leq p \leq \theta$,
   - $\eta'_\theta(p) = \frac{1}{2(1-\theta)}h'(\pi) = \frac{1}{2(1-\theta)}\left(\log_2(1-\pi) - \log_2 \pi\right)$, for $\theta \leq p \leq 1$

   The derivative is null for $\pi = 0.5$, *i.e.* $p = \theta$.
3. **Minimality**: $\eta_\theta(p)$ is minimal for $\pi = 0$ ($p = 0$) and $\pi = 1$ ($p = 1$).
4. **Concavity**: from the previous expression:
   - $\eta''_\theta(p) = \frac{1}{4\theta^2}h''(\pi) = \frac{-1}{4\theta^2 Ln2}\frac{1}{\pi(1-\pi)}$, when $0 \leq p \leq \theta$,
   - $\eta''_\theta(p) = \frac{1}{4(1-\theta)^2}h''(\pi) = \frac{-1}{4(1-\theta)^2 Ln2}\frac{1}{\pi(1-\pi)}$, when $\theta \leq p \leq 1$

   Thus, $\eta_\theta(p)$ is a concave function of p. It is to be noticed that when $p = \theta$, the left second derivative differs from the right one.

4 Off-centered entropy for a class variable having q modalities

To extend the definition of the off-centered entropy to the case of a variable Y having q modalities, $q > 2$, we follow a similar way as in the boolean case. Let $\underline{p} = (p_1, p_2, ...p_q)$ be the vector of frequencies of Y and $\underline{\theta} = (\theta_1, \theta_2, ...\theta_q)$ the reference distribution frequencies vector, for example the *a priori* distribution of Y in supervised learning. The entropy of $\underline{p}$ is $h(\underline{p}) = -\sum_{j=1}^{q} p_j \log_2 p_j$ whereas we want to express the off-centered entropy of $\underline{p}$ by $\eta(\underline{p}) = -\sum_{j=1}^{q} \pi_j \log_2 \pi_j$ where:

- $0 \leq \pi_j \leq 1$, $\sum_{j=1}^{q} \pi_j = 1$ (π_j should be analogous to frequencies)
- π_j increases from 0 to $1/q$, when p_j increases from 0 to θ_j
- π_j increases from $1/q$ to 1, when p_j increases from θ_j to 1

In [Lallich *et al.*, 2007], by looking for an expression of π_j as $\pi_j = \frac{p_j - b}{a}$, on both intervals $0 \leq p_j \leq \theta$ and $\theta \leq p_j \leq 1$, we show that :

$$\pi_j = \frac{p_j}{q\theta_j} \text{ if } 0 \leq p_j \leq \theta_j, \;\; \pi_j = \frac{q(p_j - \theta_j) + 1 - p_j}{q(1 - \theta_j)} \text{ if } \theta_j \leq p_j \leq 1$$

It is easy to check that all desirable properties are satisfied except the normalization one. We then normalize the π_j to obtain $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^{q} \pi_j}$. This normalization preserve the properties and thus we define the off-centered entropy for a variable with q modalities by $\eta_{\underline{\theta}}(\underline{p}) = h(\underline{\pi}^*)$.

5 Off-centered generalised entropies

Shannon's entropy is not the only diversity or uncertainty function usable to build coefficients of predictive association. [Goodman and Kruskal, 1954] already proposed a unified view of the three usual coefficients which are the λ of Guttman, the u of Theil and the τ of Goodman and Kruskal, under the name of *Proportional Reduction in Error* coefficient. In a more general way (*cf.* [Lallich, 2002] for details on the coefficients quoted here) we built the *Proportional Reduction in Diversity* coefficients, which are the analogue of the standardized gain when Shannon's entropy is replaced by whichever function of uncertainty. As pointed out by C. d'Aubigny ([d'Aubigny, 1980]), such a construction is justified since the function of uncertainty is concave, so the average reduction of diversity of Y with respect to X is positive, using Jensen's inequality. If the selected function I is the quadratic entropy of Gini, $I(Y) = 2(1 - \sum_{j=1}^{q} p_j^2)$ (diversity index of Gini-Simpson) the relative gain corresponds to the coefficient τ of Goodman and Kruskal, whereas if I is $I(Y) = q - 1$ (index of diversity of the number of species, in ecology) the relative gain corresponds to the coefficient λ of Guttman, Goodman and Kruskal.

More generally, we noticed that the functions of uncertainty usable were either the generalized entropies of order β of [Daroczy, 1970], or either rank diversities of order ρ introduced by [Patil and Taillie, 1982]. [Lallich, 2002] proposed a unique way of writing most of the usual coefficients in the form of a standardized reduction of generalized entropies or diversity of ranks:

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

In this formula, I refers to the entropies of order β or to their equivalent in terms of diversity of ranks of order ρ, whereas α is at the disposal of the user to choose between the two usual normalizations. This expression allows to recover the usual coefficients ($\alpha = 1$) founded on a generalized entropy ($\beta = 0$: number of categories; $\beta = 1$: Theil; $\beta = 2$: Gini) or on ranks ($\rho = 0$: Guttman; $\rho = 1$: Utton), as well as the gain-ratio ($\alpha = 0$) and Kvalseth's

coefficient ($\alpha = 0.5$). It also allows to generalize new ones. The strategy of decentring that we proposed applies without difficulties if the function of uncertainty is a generalized entropy or an entropy of ranks.

For instance, the general formula of the generalized entropies of order β is $H_\beta(\underline{p}) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - \sum_{j=1}^{q} p_j^\beta\right)$. To decentre this entropy, one first has to transform the frequencies p_j to π_j, and second to normalize these π_j, in order to obtain the pseudo-frequencies π_j^*, as described in the previous section. We obtain the distribution $\underline{p}$ by the off-centered entropy of order β, forming:

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

The off-centered versions of the entropies of ranks are built in the same way. For instance, $H_{\rho=0}(\underline{p}) = 2\left(1 - \max\left\{p_j, j = 1, 2, ..., q\right\}\right)$, thus:

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

Figure 2 well illustrates the behaviour of the off-centered generalized entropies we propose ($\beta = 0, 0.5, 1, 2, 5$, $\rho = 0$ and the asymmetrical entropy proposed by [Zighed *et al.*, 2007]) where the *a priori* distribution of the class variable is $(0.8, 0.2)$ that corresponds to $\theta = 0.2$. We propose a decentring framework that one can apply to any measure of predictive association based on a gain of uncertainty. The choice of the value of β or ρ depends on the reactivity one expects from the measure.

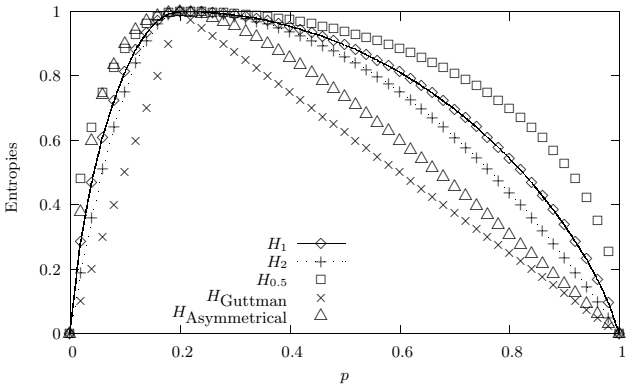

Fig. 2. Off-centering of generalised entropies

6 Conclusion and future works

Usual predictive association measures may be expressed in terms of normalised gains, which is associated to an uncertainty function, generalised

entropy or diversity of ranks. We proposed an off-centering methodology associating an entropy with whichever generalised previous expressions (especially β order entropies and ρ orders liking-ranks).

Further investigations should sure be carried out on data, and thus show the interest of such an approach, especially when the class variable is quite imbalanced. The predictive performances of supervised learning algorithms should hence be enhanced in that way.

References

[Daroczy, 1970]A. Daroczy. Generalized information functions. *Information and Control*, (16):36–51, 1970.

[d'Aubigny, 1980]C. d'Aubigny. *Etude de la morphologie des indices d'association*. PhD thesis, Université Joseph Fourier, Grenoble, France, 1980.

[Goodman and Kruskal, 1954]L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications, i. *JASA*, I(49):732–764, 1954.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *EGC 2001*, volume 1, pages 69–80, Nantes, France, 2001.

[Kvalseth, 1987]T. O. Kvalseth. Entropy and correlation: some comments. *IEEE Trans. on Systems, Man and Cybernetics*, 17(3):517–519, 1987.

[Lallich *et al.*, 2005]S. Lallich, B. Vaillant, and P. Lenca. Parametrised measures for the evaluation of association rule interestingness. In *ASMDA 2005*, pages 220–229, Brest, France, 2005.

[Lallich *et al.*, 2007]S. Lallich, B. Vaillant, and P. Lenca. Construction d'une entropie décentrée pour l'apprentissage supervisé. In *QDC/EGC 2007*, pages 45–54, Namur, Belgium, 2007.

[Lallich, 2002]S. Lallich. Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches, Université Lyon 2, France, 2002.

[Marcellin *et al.*, 2006]S. Marcellin, D.A. Zighed, and G. Ritschard. An asymmetric entropy measure for decision trees. In *IPMU 2006*, pages 1292–1299, Paris, France, 2006.

[Patil and Taillie, 1982]G.P. Patil and C. Taillie. Diversity as a concept and its measurement. *J. of American Statistical Association*, 77(379):548–567, 1982.

[Quinlan, 1975]J.R. Quinlan. *Machine Learning*, volume 1. 1975.

[Quinlan, 1993]J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[Shannon, 1948]C. E. Shannon. A mathematical theory of communication. *Bell System Technological Journal*, (27):379423, 623656, July and October 1948.

[Theil, 1970]H. Theil. On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, (76):103–154, 1970.

[Wehenkel, 1996]L. Wehenkel. On uncertainty measures used for decision tree induction. In *IPMU 1996*, pages 413–418, 1996.

[Zighed and Rakotomalala, 1998]D. A. Zighed and R. Rakotomalala. *Graphes d'induction et apprentissage machine*. Hermès Paris, 1998.

[Zighed *et al.*, 2007]D.A. Zighed, S. Marcellin, and G. Ritschard. Mesure d'entropie asymétrique et consistante. In *EGC 2007*, pages 81–86, Namur, Belgium, 2007.