L'usage des entropies est-il justifie en apprentissage a partir des donnees?

Is it justified to use entropy measures in machine learning applications?

DJAMEL A. ZIGHED[1]

Résumé
De nombreux algorithmes d'apprentissage machine utilisent les mesures d'entropie comme critère de construction qu'ils cherchent ensuite à optimiser. Parmi les mesures le plus employées, l'entropie de Shannon est certainement la plus populaire. Cependant, dans les applications réelles, l'usage des mesures d'entropie s'avère totalement inapproprié à la fois sur le plan pratique et sur le plan théorique. De nombreuses hypothèses sont en fait retenues de manière implicites alors qu'elles sont infondées. Dans cette présentation, nous allons essayer d'identifier ces hypothèses sous-jacentes et montrer qu'elles sont inadaptées en apprentissage à partir des données. Nous énoncerons ensuite, de façon intuitive d'abord, de nouvelles propriétés qui se requises pour définir des mesures pouvant déboucher sur des algorithmes plus efficients pour l'apprentissage machine.
Mots-clés: *Mesures d'entropie, Apprentissage machine*

Abstract
Many machine learning algorithms use entropy measures as a criterion of construction that they seek to optimize. Among the most applied measures, Shannon's entropy is certainly the most known. However, in the real world applications, the use of the entropy measure turns out to be totally inadequate both in theory and in practice. Indeed, many hypothesis are in fact implicitly assumed whereas they are unfounded, therefore unjustified. In this paper, we will try to identify those hypothesis and we will demonstrate that they are unsuitable in machine learning with real data. Then, we will introduce, intuitively, a set of new prosperities that should be required for measures that are supposed to lead to efficients algorithms.
Keywords: *Entropy measures, Machine learning*

Introduction

En apprentissage machine, plus particulièrement en apprentissage supervisé, de nombreuses classes d'algorithmes, comme ceux basés sur les règles d'association ou bien les arbres de décision, font appel à des mesures d'entropie. Ces mesures sont appliquées sans vraiment bien expliciter les hypothèses qui justifieraient leur usage. En effet, ces mesures, qui ont été introduites à l'origine dans des contextes autres que l'apprentissage supervisé, reposent sur des propriétés qui sont rarement vérifiées dans les applications pratiques de l'apprentissage automatique. Pour contourner les difficultés et les limitation

[1] Université Lumière Lyon 2, Laboratoire ERIC, Bât L, Campus Porte des Alpes 5, av. Pierre Mendès-France, F-69600 Bron, France abdelkader.zighed@ish-lyon.cnrs.fr http://zighed.com

et effet indésirables des entropies, les auteurs ont introduit dans leurs algorithmes de nombreux paramètres d'ajustement guidés surtout par une intuition que par une critique de fond. Dans ce travail nous allons montrer cette inadéquation pour proposer à la fois une définition d'une nouvelle axiomatique pour des critères d'apprentissage et une mesure particulière qui réponde à ces nouvelles exigences. Dans la section deux, nous donnons les principales notations. Dans la section trois nous allons d'abord rappeler l'axiomatique des mesures d'entropie pour donner ensuite quelques exemples de mesures. Dans la section quatre, nous allons passer en revue ces axiomes et montrer en quoi ils ne satisfont pas à la plupart des conditions que l'on trouve en apprentissage supervisé. Nous terminons en section cinq par une proposition, sans détailler les démonstrations, d'une nouvelle mesure qui vérifie les nouvelles exigences requises.

Notations et définitions

Dans la suite, on note Ω la population concernée par la question d'apprentissage. A tout individu ω dans Ω de cette population, généralement infinie, sont observées les états de p variables dites descriptives ou prédictives notées $X_1, \dots, X_p$. Ces variables peuvent être quantitatives ou qualitatives. On considère également une variable C ayant un intérêt particulier. Il s'agit généralement d'une variable qualitative à m modalités notées $c_j, j = 1, \dots, m$ que l'usager cherche à prédire au moyen des variables prédictives. De façon générale, les algorithmes d'apprentissage tentent d'expliciter un modèle ϕ combinant, tout ou partie, des variables $X_j; j = 1, \dots, p$ pour prédire la valeur de C. C'est le cas pour les arbres de décision [17, 23], des graphes d'induction [15] et de l'ensemble des méthodes à base de règles qui font appel à des mesures d'entropie. En fait, l'usage des entropies est possible dès lors que nous pouvons induire une partition notée S sur l'ensemble d'apprentissage. Chaque élément s de cette partition sera ensuite caractérisé par une distribution de fréquences relatives aux différentes classes d'apprentissages c_j. Cela permet ensuite de calculer une entropie en assimilant les fréquences $f(c_j/s)$ aux probabilités des classes comme on va le voir ci-après. Dans le cas général d'une partition à plusieurs éléments, l'entropie de la partition sera la moyenne pondérée des entropies des différents éléments de la partition. Les partitions sont engendrées par les combinaisons des différentes variables X_j. Chaque élément étant défini par une règle sous forme d'une disjonction de conjonctions.

Pour illustrer notre propos, nous reprenons les données fictives utilisées dans [15] et récapitulées au tableau 1. La population Ω est celle des ménages dont on a extrait un échantillon de 273 individus. La variable C à prédire est l'état civil, le sexe et le secteur d'activité étant les prédicteurs disponibles $X_j; j = 1,2$.

Tableau 1. Exemple illustratif

	Hommes			Femmes			Total
	Primaire	Secondaire	Tertiaire	Primaire	Secondaire	Tertiaire	
Marié	50	40	6	0	14	10	**120**
Célibataires	5	5	12	50	30	18	**120**
Divorcé/Veuf	5	8	10	6	2	2	**33**
Total	**60**	**53**	**28**	**56**	**46**	**30**	**273**

Mesures d'entropie

Le concept d'entropie a été introduit par Hartley [8] mais a été réellement développé et utilisé notamment dans les télécommunications par Shannon et Weaver dans les années 40 [13, 14]. Ces deux auteurs ont proposé une mesure d'information qui est en fait une entropie sur une distribution de probabilités. Ces travaux ont été poursuivis et approfondis par de nombreux autres chercheurs comme Hencin [9] plus tard, Forte [7], Aczel et Daroczy [1] qui ont fondé l'axiomatique des mesures d'entropie dans le contexte de la théorie de l'information. Parmi les formules célèbres et couramment employées, figure l'entropie de Shannon.

Entropie de Shannon

Soit E une expérience pouvant produire les événements $e_1, e_2, \dots, e_m$ ayant chacun une probabilité associée $p_1, p_2, \dots, p_m$. On suppose donc que $\sum_i^m p_i = 1$ et $p_i \geq 0$ pour $i = 1, \dots, m$. L'entropie de Shannon d'une distribution de probabilité sur m événements discrets est donnée par la formule :

$$H(p_1, p_2, \dots, p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (1)$$

Par continuité, on pose $0\log_2 0 = 0$.

Entropie sur une partition

Si sur une population E nous engendrons une partition S d'éléments génériques s disjoints et non vides. Sur chaque élément s nous supposons connaitre $P(s)$ (probabilité a priori

de s) et $P(c_i/s); i = 1, \dots, m)$ (probabilités conditionnelles des classes), alors l'entropie de la partition S est :

$$H(S) = \sum_{s \in S} P(s) H(P(c_1|s), \dots, P(c_i|s), \dots, P(c_m|s)) \quad (2)$$

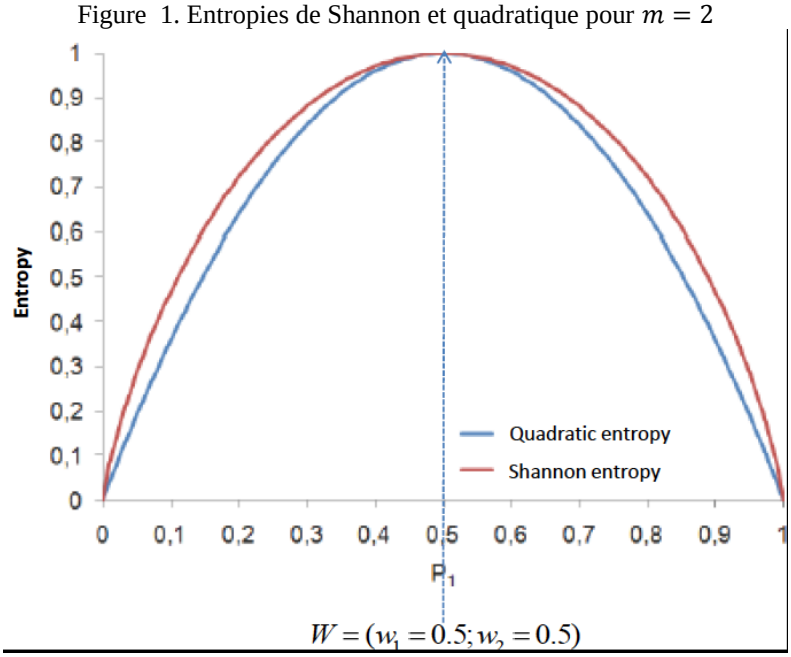
Quelques exemples de mesures d'entropie

Il y a de nombreuses formules d'entropies disponibles dans la littérature scientifique [11] [16] dont l'une des plus connue est celle basée sur l'indice de concentration de Gini et souvent appelée entropie quadratique :

$$H(p_1, p_2, \dots, p_m) = \sum_{i=1}^{m} p_i(1 - p_i) \quad (3)$$

La figure 1 représente les entropies de Shannon et quadratique pour une distribution à deux classes ($m = 2$).

Figure 1. Entropies de Shannon et quadratique pour $m = 2$



Définition formelle et propriétés communes aux mesures d'entropie

Pour Soit une distribution de probabilité sur un espace discret $(p_1, p_2, \dots, p_m)$ avec $m \geq 2$. On notera Γ_m le simplexe d'ordre m:

$$\Gamma_m = \{(p_1, p_2, \dots, p_m): \sum_i^m p_i = 1; p_i \geq 0\} \quad (4)$$

De façon formelle, une mesure d'entropie est définie comme suit :

$$h: \Gamma_n \to R \quad (5)$$

vérifiant les propriétés suivantes :

A1 : Non négativité

$$h(p_1, p_2, \dots, p_m) \geq 0 \quad (6)$$

A2 : Symétrie

L'entropie est indépendante de toutes permutations du vecteur $(p_1, \dots, p_m)$ de Γ_m.

$$(A2) : h(p_1, p_2, \dots, p_m) = h(p_{\sigma_{(1)}}, p_{\sigma_{(2)}}, \dots, p_{\sigma_{(m)}}) \quad (7)$$

où σ est une permutation quelconque de $(p_1, p_2, \dots, p_m)$.

A3 : Minimalité

s'il existe un k tel que $p_k = 1$ et que $p_i = 0 \forall i \neq k$ alors

$$h(p_1, p_2, \dots, p_m) = 0 \quad (8)$$

A4 : Maximalité

$$h(p_1, p_2, \dots, p_m) \leq h(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}) \quad (9)$$

A5 : Concavité stricte

La fonction $h(p_1, p_2, \dots, p_m)$ est strictement concave.

Sur l'inadéquation des entropies en apprentissage supervisé

Dans la suite, nous allons passer en revue les principales priorités énoncées plus haut et montrer qu'elle sont très rarement satisfaites :

- **La calculabilité effective de l'entropie**

L'usage direct d'une entropie sans autre facteur de correction pose un vrai problème de qualité du modèle. En effet, les probabilités des classes sont estimées par les fréquences empiriques sur l'échantillon d'apprentissage. Cela suppose que cet échantillon soit suffisamment grand pour que ces estimations aient un sens sur le plan statistique. Plus encore, cela suppose que chaque élément s d'une quelconque partition S sur l'échantillon d'apprentissage E ait lui-même un cardinal suffisant pour que les estimations $P((.)/s); \forall s \in S$ soient crédibles pour l'usager. Or dans les approches à base de règles, comme les arbres de décision, le processus se déroule par raffinement itératif de la partition grossière vers des partitions de plus en plus fines dont les éléments sont de faible entropie. Pour tenir compte de ce problème, beaucoup d'algorithmes proposent de fixer un seuil minimal pour tout élément s d'une partition. Une sorte de seuil d'admissibilité

pour autoriser ou non un raffinement donné. Cette solution est acceptable mais n'est pas satisfaisante car elle introduit un paramètre en dehors du critère d'entropie lui-même.

• **Tendance au sur-apprentissage**

On peut démontrer que si S' est une partition plus fine issue de S alors

$$H(S') \leq H(S)$$

Cela signifie que l'entropie moyenne ne croît jamais avec le raffinement de la partition. Autrement dit encore, la partition de plus faible entropie est la partition la plus fine que l'on peut engendrer par les variables. Dans un arbre de décision par exemple, la partition d'entropie minimale serait celle qui résulterait de l'arbre le plus grand possible ce qui nous ramène au problème de la calculabilité effective de l'entropie précédent. Trop de noeuds entraine une dispersion de l'information avec des effectifs faibles en chaque noeud. La partition est certes de faible entropie mais les règles induites se trouvent alors trop spécialisées sur quelques individus et de ce fait auront une faible capacité à se généraliser à toute la population: phénomène connu de sur-apprentissage. C'est cet inconvénient qui a conduit [21] à introduire le processus d'élagage dans les arbres pour tenter de retrouver la partition intermédiaire à laquelle il aurait dû s'arrêter si l'entropie était sensible à cette question. Plus encore, dès lors que cette propriété de raffinement qui garantit la décroissance de l'entropie, on pourrait même se passer d'utiliser une entropie dans la construction des règles. Ce point a été soulevé au début du développement des arbres de décision. L'usage de l'entropie servirait juste alors comme indicateur sur l'importance relative des variables.

• **La non prise en compte de la complexité du modèle**

La référence à la partition la plus fine comme étant la meilleure au regard de l'entropie va nécessairement conduire à des modèles de plus en plus complexes. C'est-à-dire, des modèles avec beaucoup de règles allant ainsi dans le sens contraire du grand principe du rasoir d'occam [18]. Encore ici, Breiman [21] a introduit dans son critère d'élagage cette notion de complexité mesurée par le nombre de noeuds fils.

• **La référence à la distribution uniforme comme la plus mauvaise**

Dans les propriétés des entropies montrées plus haut, la distribution uniforme est celle dont l'entropie est maximum. Tous les algorithmes qui utilisent une entropie comme critère à optimiser vont nécessairement chercher à s'éloigner en moyenne de la

distribution uniforme dans les éléments de la partition S à construire. Or dans la pratique, la distribution uniforme n'est pas nécessairement la plus mauvaise aux yeux de l'usager. En effet, l'usager chercherait plutôt à s'éloigner de la distribution a priori des classes $P(c_i)$. Prenons l'exemple d'un problème d'identification de fraudes où la proportion d'incidents "fraude" est infinitesimale par rapport à la classe opposée. Par conséquent une situation avec une distribution uniforme sur les deux classes est particulièrement significative pour l'usager par rapport à la situation initiale. Pourtant un pur calcul de l'entropie lui indiquerait l'inverse. Pour tenir compte de cet inconvénient, des auteurs comme [24] prônent le rééquilibrage des classes par rééchantillonnage dans la classe sur représentée.

- **L'hypothèse de symétrie**

La propriété de symétrie est rarement vérifiée dans les problèmes d'apprentissage supervisé. Au yeux du décideur, souvent les classes n'ont pas la même importance. Pour garder l'exemple de détection des fraudes cité plus haut, aux yeux de l'usager, l'information tirée d'une distribution avec 90% de fraudes n'est pas équivalente à celle d'une distribution avec 10% de fraude. pourtant, l'entropie associée à cette distribution sera la même. Pour contrecarrer cet effet, des auteurs [19, 20] ont proposé de rendre asymétrique la mesure d'entropie. Mais les propositions se heurtent à des difficultés d'interprétation de la généralisation du principe à m classes.

Discussion et conclusion

L On voit clairement que l'utilisation directe et sans autres éléments de correction d'une entropie, quelle qu'elle soit, n'apporte pratiquement rien. Tous les auteurs [2], [3], [4], [5], [6], [10], [22], [12], [15] ont tenté intuitivement de compenser les insuffisances mentionnées plus haut en introduisant des modification dans le critère finale, en rajoutant des conditions ad hoc ou en développant des algorithmes complémentaires pour corriger les distorsions produites par l'entropie. Dans notre travail [12] nous avons essayé de refondre une axiomatique visant à intégrer l'ensemble des critiques mentionnées. Nous avons alors proposé une famille de mesure d'entropie à valeur non négative:

$$h_W(N, f_1, f_2, \dots, f_m) = \sum_{i=1}^{m} \frac{\lambda_i(1-\lambda_i)}{(-2w_i+1)\lambda_i + w_i^2}$$

Avec

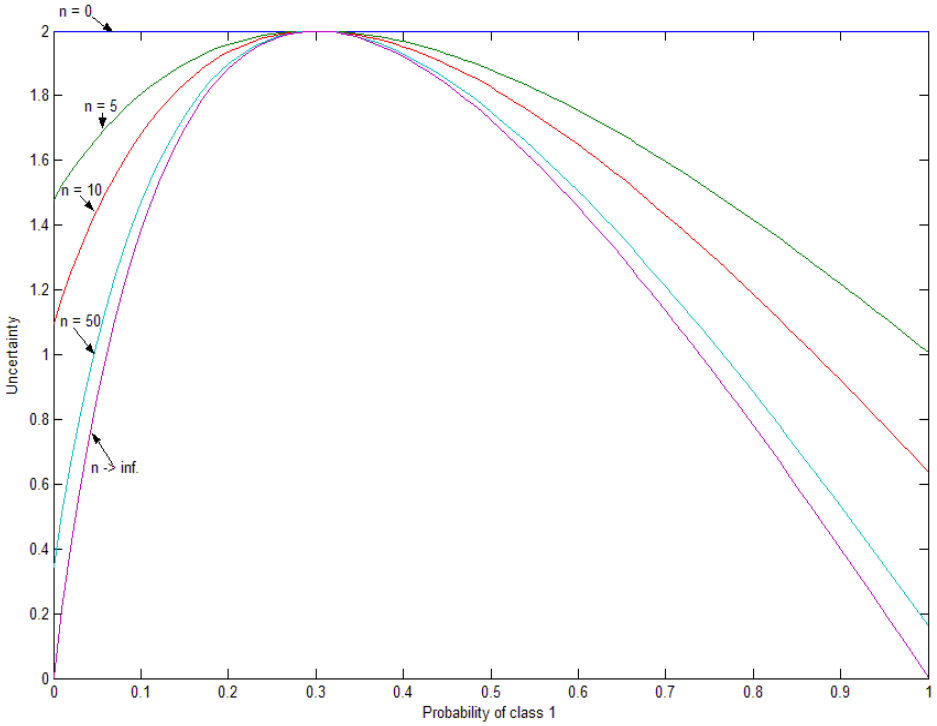
• N la taille de l'échantillon d'apprentissage

• $\lambda_i = \frac{Nf_i+1}{N+m}$ l'estimateur de Laplace des probabilités

• $W = (w_1, w_2, \ldots, w_m)$ la distribution a priori des classes estimée sur l'échantillon d'apprentissage.

Pour un exemple à deux classes, si nous fixons la distribution a priori à $f_1 = 0.3$ et $f_2 = 0.7$ nous pouvons tracer la courbe de l'entropie en fonction de la valeur de N, taille de l'échantillon comme cela est montré sur la figure 2.

Figure 2. Mesure d'entropie asymétrique et sensible aux effectifs



Cette mesure vérifie une liste de propriétés que nous énumérons ci-dessous mais sans en donner les démonstrations.

• sensible aux effectifs,

• asymétrique,

• prenant son maximum pour la distribution initiale

• prenant en compte la complexité

• qui toujours décroissante

Références

ACZEL, J., Daroczy, Z.: On Measures of Information and Their Characterizations. Academic Press, NY, S. Francisco, London (1975)

BARANDELA, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition **36(3)** (2003) 849–851

CHAI, X., Deng, L., Yang, Q., Ling: Test-cost sensitive naive bayes classification. In IEEE, ed.: ICDM apos;04. Fourth IEEE International Conference on Data Mining, ICDM'04 (2004) 973–978

CHEN, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Technical Report 666, Berkeley, Department of Statistics, University of California (2004)

DOMINGOS, P.: Metacost: A general method for making classifiers cost-sensitive. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99) (1999) 155–164

ELKAN, C.: The foundations of cost-sensitive learning. In Nebel, B., ed.: IJCAI, Morgan Kaufmann (2001) 973–978

FORTE, B.: Why shannon's entropy. In Conv. Inform. Teor., **15** (1973) 137–152

HARTLEY, R.V.: Transmission of information. Bell System Tech. J. **7** (1928) 535–563

HENCIN, A.J.: The concept of entropy in the theory of probability. Math. Found. of Information Theory (1957) 1–28

PROVOST, F.: Learning with imbalanced data sets. Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets (2000)

RÉNYI, A.: On measures of entropy and information. 4th Berkely Symp. Math. Statist. Probability **1** (1960) 547–561

RITSCHARD, G., Zighed, D., Marcellin, S.: Données déséquilibrées, entropie décentrée et indice d'implication. In Gras, R., Orús, P., Pinaud, B., Gregori, P., eds.: Nouveaux apports théoriques à l'analyse statistique implicative et applications (actes des 4èmes rencontres ASI4, 18-21 octobre 2007), Castellón de la Plana (España), Departament de Matemàtiques, Universitat Jaume I (2007) 315–327

SHANNON, C.E.: A mathematical theory of communication. Bell System Tech. J. **27** (1948) 379–423

SHANNON, C.A., Weaver, W.: The mathematical of communication. University of Illinois Press (1949)

ZIGHED, D.A., Marcellin, S., Ritschard, G.: Mesure d'entropie asymétrique et consistante. In Noirhomme-Fraiture, M., Venturini, G., eds.: EGC. Volume RNTI-E-9 of Revue des Nouvelles Technologies de l'Information., Cépaduès-Éditions (2007) 81–86

ZIGHED, D., Rakotomalala, R.: Graphe d'induction: Apprentissage et Data Mining. Hermès, Paris (2000)

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

DOMINGOS, P.: The role of occam's razor in knowledge discovery. Data mining and knowledge discovery **3**(4) (1999) 409–425

LENCA, P., Lallich, S., Do, T.N., Pham, N.K.: A comparison of different off-centered entropies to deal with class imbalance for decision trees. In: Advances in Knowledge Discovery and Data Mining. Springer (2008) 634–643

MARCELLIN, S., Zighed, D.A., Ritschard, G.: Evaluating decision trees grown with asymmetric entropies. In: Foundations of Intelligent Systems. Springer (2008) 58–67

OLSHEN, L.B.J.F.R., Stone, C.J.: Classification and regression trees. Wadsworth International Group (1984)

PROVOST, F.J., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. Knowledge Discovery and Data Mining (1997) 43–48

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

SEBBANÜ, M., NockO, R., Chauchat, J., Rakotomalala, R.: Impact of learning set quality and size on decision tree performances. IJCSS **1**(1) (2000) 85