Asymmetric and sample size sensitive entropy measures for supervised learning

Zighed, Djamel A.; Ritschard, Gilbert

Asymmetric and sample size sensitive entropy measures for supervised learning

Djamel A. Zighed and Gilbert Ritschard and Simon Marcellin

Abstract Many algorithms of machine learning use an entropy measure as optimization criterion. Among the widely used entropy measures, Shannon's is one of the most popular. In some real world applications, the use of such entropy measures without precautions, could lead to inconsistent results. Indeed, the measures of entropy are built upon some assumptions which are not fulfilled in many real cases. For instance, in supervised learning such as decision trees, the classification cost of the classes is not explicitly taken into account in the tree growing process. Thus, the misclassification costs are assumed to be the same for all classes. In the case where those costs are not equal on all classes, the maximum of entropy must be elsewhere than on the uniform probability distribution. Also, when the classes don't have the same a priori distribution of probability, the worst case (maximum of the entropy) must be elsewhere than on the uniform distribution. In this paper, starting from real world problems, we will show that classical entropy measures are not suitable for building a predictive model. Then, we examine the main axioms that define an entropy and discuss their inadequacy in machine learning. This we lead us to propose a new entropy measure that possesses more suitable proprieties. After what, we carry out some evaluations on data sets that illustrate the performance of the new measure of entropy.

ERIC Lab. University of Lyon 2
5, av. Pierre Mends-France, 69600 Bron, France, e-mail: abdelkader.zighed@univ-lyon2.fr

University of Geneva, Dep. of econometry
bd du Pont-d'Arve, H-1211 Geneva 4, Switzerland e-mail: gilbert.ritschard@unige.ch

ERIC Lab. University of Lyon 2
5, av. Pierre Mends-France, 69600 Bron, France, e-mail: simon.marcellin@univ-lyon2.fr

1 Introduction

In machine learning, more specifically in supervised learning, algorithms such as association rules, decision trees,... use plenty of criteria, and among them are measures of entropy. Unfortunately, when the entropy criteria are used, it is done without taking into account the assumptions upon which they are founded. Indeed, many assumptions required for such usage are not satisfied in real applications. The entropy criteria would be suitable if, in one hand, the classes were balanced, i.e. they had, almost, the same a priori probability and, on the other hand, the misclassification costs were equal for all the classes. Entropy measures are also based on an axiomatic which assumes that the probabilities of the classes could be calculated at any time, which is not always possible because of the finite size of the learning sample. Let us describe some situations when the main assumptions are not taken into consideration:

- **Hypothesis of distribution of classes a priori uniform** : This hypothesis is not valid in real world applications. We can observe this when the classes are unbalanced. In such case, the distribution of the modalities of the class variable is far away from the uniform distribution. If the sampling process does not suffer from any bias, i.e. the sample conforms to the reality, then we may conclude that the a priori distribution of the classes is not uniform. This happens in a lot of real world applications: in the medical field, to predict a rare illness; in the industry to predict a device failure; or in the banking field, to predict insolvent customers or frauds in transactions. In these cases, there is one rare state of the class variable (ill, breakdown, insolvent, fraud) with less cases in comparison to the whole population. Standard methods do not take such specificities into account and just optimize a global criterion with the consequence that all the examples would be classified into the majority class, i.e. which minimizes the global error rate on the learning set. This kind of prediction models is useless because it does not carry any information. In decision trees, this problem appears at two levels: during the generation of the tree with the splitting criterion, and during the prediction with the assignment rule of a class in each leaf. Indeed, in decision tree for instance, to choose the best feature and the best split point to create a new partition, classical algorithms use an entropy measure, like the Shannon entropy [23] and [22] or quadratic entropy [26]. Entropy measures evaluate the quantity of information about the outcome provided by the distribution of the class variable. They consider the uniform distribution, i.e for which we have the same number of cases in each class, as the most entropic situation. So the worst situation according to these measures is the balanced distribution. However, if in the real world for example a priori 1% of the people are sick, ending with a leaf in which 50% of the members are sick would be very instructive and would carry a lot of information for the user. Thus, using a classical entropy measure precludes obtaining such branches and hence the relevant associated rules for predicting the rare class. The second important aspect of decision trees is the assignment rule. Once the decision tree is grown, each branch defines the condition of a rule. The conclusion

of the rule depends on the distribution of the leaf. Classical algorithms conclude to the majority class, i.e the most frequent modality in the leaf. But this is not efficient: In the previous example where 1% of the people are sick, a rule leading to a leaf with a frequency of the 'sick' class of 30% would conclude to 'not sick'. According to the importance of predicting correctly the minority class, it may be better however in that case to conclude to 'sick'. This will lead to a higher total number of errors, but a lower number of errors on the rare class and hence a better model.

- **Hypothesis of equal misclassification costs** : Overall, the supervised learning algorithms assume that the misclassification costs are equal for all the classes, thus the cost is constant and fixed. If we denote by c_{ij} the cost of the classification of an individual issued from the class i to the class j then, we have :
  - a symmetrical misclassification cost : $c_{ij} = c_{ji} = c$ for all $(i, j); i \neq j$
  - the cost of a good classification $c_{ii} = 0$ for all classes

  But, in many real world applications, this hypothesis is not true. For instance, in cancer diagnosis, missing a cancer could lead to death whereas the consequence of misleading to a cancer are less important even if they are costly.
- **Hypothesis of non sensitivity to the sample size** : the entropy measures are all non sensitive to the sample size. They depend only on the distribution of the classes. For instance, in decision trees, if we consider two leaves, with the same distribution of the classes, the values of the entropy associated to each node are equal even if one node has many more individuals. Yet it would be natural to consider the leaf, with the higher size, as providing a more reliable information.

Plenty of works have been done to address issues brought about by the above assumptions. We may cite [2], [4], [5], [6], [8], [16], [17], [20], [25]. All these works have dealt with only one issue at a time. In section 2 we introduce some notations and definitions. We will focus on the axiomatic of the entropy which has been defined, at the beginning, outside of the area of machine learning and then we will present some measures of entropy. In section 3 we introduce our design for a new entropy measure that fulfill a set of requirements. In section 4 we propose an evaluation based on some experiments on data set where some are drawn from real world applications. And then, in section 5, we conclude and propose some new directions.

2 Notations and basic definition

For the sake of clarity of the presentation, our frame work is that of decision trees. Nevertheless, our proposal may be extended to any other machine learning algorithms that use entropy measure as criterion.

2.1 Notations and basic concepts

We denote Ω the population concerned by the learning problem. The profile of any example ω in Ω is described by p explicative or exogenous features $X_1, \ldots, X_p$. Those features may be qualitative or quantitative ones. We also consider a variable C to be predicted called either endogenous, class or response variable. The values taken by this variable within the population are discrete and form a finite set $\mathscr{C}$. Letting m_j be the number of different values taken by X_j and n the number of modalities of C, we have $\mathscr{C} = \{c_1, \ldots, c_n\}$. And when it is not ambiguous, we denote the class c_i simply by i. Algorithms of tree induction generate a model $\phi(X_1, \ldots, X_p)$ for the prediction of C represented by a decision tree [3, 18] or an induction graph [25]. Each branch of the tree represents a rule. The set of these rules is the prediction model that permits to determine the predicted value of the endogenous variable for any new example for which we only know the exogenous features. The development of the tree is made as follows: The learning set Ω_a is iteratively segmented, each time on one of the exogenous features $X_j; j = 1, ...p$ so as to get the partition with the smallest entropy for the distribution of C. The nodes obtained at each iteration define a partition on Ω_a. Each node s of a partition S is described by a probability distribution of the modalities of the endogenous features C: $p(i/s); i = 1, \ldots, n$. Finally, these methods generate decision rules in the form **If** *condition* **then** *Conclusion*. Splitting criteria are often based on entropies.

2.2 Entropy measures

The concept of entropy has been introduced by Hartley [11] but was really developed and used in the industrial context by Shannon and Weaver [22, 23] in the forties. They proposed a measure of information which is the general entropy of a distribution of probabilities. Following the theorem that defines the entropy, many researchers such as Hencin [12] and later, Forte [10], Aczel and Daroczy [1] have proposed an axiomatic approach for the entropies.

2.2.1 Shannon's entropy

Let E be an experience with the possible events $e_1, e_2, \ldots, e_n$ of respective probabilities $p_1, p_2, \ldots, p_n$. We suppose that $\sum_i^n p_i = 1$ et $p_i \geq 0$ for $i = 1, \ldots, n$. The entropy of Shannon of the probabilities distribution is given by the following formula :

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

By continuity, we set $0 \log_2 0 = 0$.

2.2.2 Entropy on a partition

The entropy H on the partition S to minimize is generally a mean entropy such that $H(S) = \sum_{s \in S} p(s)h(p(1|s), \ldots, p(i|s), \ldots, p(n|s))$ where $p(s)$ is the proportion of cases in the node s and $h(p(1|s), \ldots, p(n|s))$ an entropy function such as Shannon's entropy for instance $H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$ ..

There are many other entropy measures [19] [26] such as the quadratic entropy $H_n(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i(1 - p_i)$ for instance. The Figure 1 depicts the quadratic and Shannon entropies for 2 classes. All the pictures of entropy measures have the same shape.

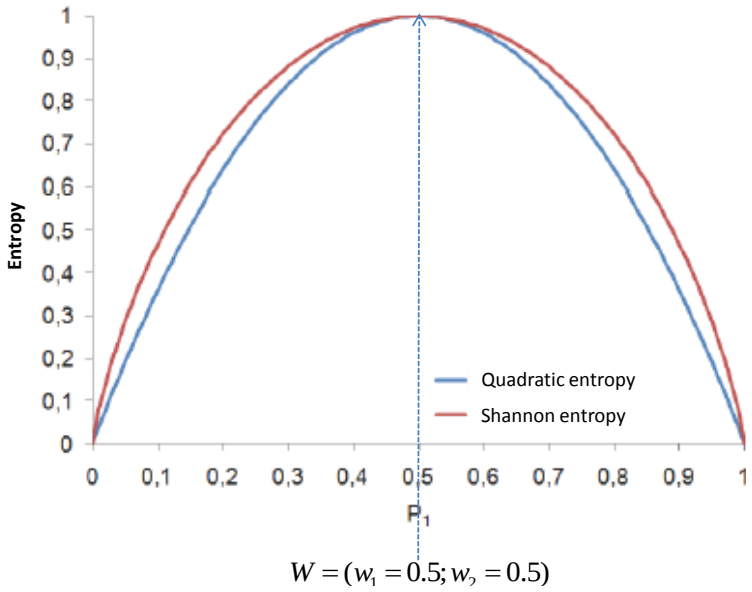

Fig. 1 Shannon and Quadratic entropies for a 2 classes problem

2.2.3 Properties of the entropy measures

Let's suppose that $(p_1, p_2, \ldots, p_n)$ for $n \geq 2$ are taken in a finite set of distributions of probabilities and let's consider the simplex of order n

$$\Gamma_n = \{(p_1, p_2, \ldots, p_n) : \sum_{i}^{n} p_i = 1; p_i \geq 0\} \tag{2}$$

A measure of entropy is defined as follow :

$$h : \Gamma_n \rightarrow \mathbf{R} \tag{3}$$

with the following properties :

Non negativity:

$$h(p_1, p_2, \ldots, p_n) \geq 0 \tag{4}$$

Symmetry: The entropy is non sensitive to any permutation within the vector $(p_1, \ldots, p_n)$ in Γ_n.

$$h(p_1, p_2, \ldots, p_n) = h(p_{\sigma_{(1)}}, p_{\sigma_{(2)}}, \ldots, p_{\sigma_{(n)}}) \tag{5}$$

where σ is any permutation on $(p_1, p_2, \ldots, p_n)$.

Minimality: If exists k such that $p_k = 1$ and that $p_i = 0$ for all $i \neq k$ then

$$h(p_1, p_2, \ldots, p_n) = 0 \tag{6}$$

Maximality:

$$h(p_1, p_2, \ldots, p_n) \leq h(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \tag{7}$$

Strict concavity: The function $h(p_1, p_2, \ldots, p_n)$ is strictly concave.

3 Asymmetric and sample size sensitive entropy

3.1 Asymmetric criteria

The properties of classical entropy measures such as those cited above (Shannon, quadratic) are not suited to inductive learning for many reasons [25]:

- First, the uniform distribution is not necessarily the most uncertain.
- Second, the computation of the entropy being based on estimates of the probabilities should account for the precision of those estimates, i.e. account for the sample size.

That is why we proposed in [25] a new axiomatic leading to a new family of more general measures. They make it possible for the user to define a reference distribution that is viewed as of maximal entropy. It permits also to make the entropy measure sensitive to the sample size.

We recall below the new axiomatic that take into account the limitations we have identified.

3.2 Properties requested for the new entropy measure

Let $\hbar$ be the new function of entropy that we want to build. We want it to be empirical, i.e. frequency dependent $f(i/.)$, sensitive to the sample size N and parametrized

by a distribution of frequencies $W = (w_1, \ldots, w_j, \ldots, w_p)$ which is considered as the less desired, i.e. where the entropy must be maximal.

$$\hbar : \mathbf{N}^* \times \Gamma_n^2 \to \mathbf{R}^+ \tag{8}$$

For a fixed distribution W, that we explain later on how it is set up, the function $\hbar_W(N, f_1, \ldots, f_i, \ldots, f_n)$ must have the following properties :

(P1) **Non negativity:** The function $\hbar$ must be non negative

$$\hbar_W(N, f_1, \ldots, f_j, \ldots, f_n) \geq 0 \tag{9}$$

(P2) **Maximality:** Let $W = (w_1, w_2, \ldots, w_n)$ be a distribution fixed by the user as the less desired and therefore of maximal entropy value. Thus, for a given N,

$$\hbar_W(N, f_1, \ldots, f_n) \leq \hbar_W(N, w_1, \ldots, w_n) \tag{10}$$

for all distribution $(f_1, \ldots, f_n)$ brought from a sample of size N.

(P3) **Asymmetry:** The new property of maximality challenges the axiom of symmetry required by the classical entropies. Therefore, some permutations σ could affect the value of the entropy : $\hbar(f_1, \ldots, f_n) \neq \hbar(f_{\sigma_1}, \ldots, f_{\sigma_n})$.
We can easily identify the conditions in which the property of symmetry would be kept. For instance in the case where w_i would be equal, i.e. in the case of uniform distribution.

(P4) **Minimality:** In the context of classical entropy, the value of the entropy is null when the distribution of the sample over the classes is concentrated in one class, in other word, it exists j such that $p_j = 1$ and that $p_i = 0$ for all $i \neq j$. This property must remain theoretically valid. However, in real world problems of supervised learning these probabilities are unknown and must be estimated.
It would still be embarrassing to say that the entropy is null when the distribution is concentrated in one specific class. We have to take into consideration the size of the sample on which the probabilities p_j are estimated.
So, we merely require that the entropy of an empirical distribution for which it exists j such that $f_j = 1$, to tend to zero when N becomes big :

$$\lim_{N \to \infty} \hbar_W(N, 0, \ldots, 0, 1, 0 \ldots, 0) = 0 \tag{11}$$

(P5) **Consistency:** For a given W and a constant distribution, the entropy must be smaller when the size of the sample is bigger.

$$\hbar_W(N+1, f_1, \ldots, f_j, \ldots, f_n) \leq \hbar_W(N, f_1, \ldots, f_j, \ldots, f_n) \tag{12}$$

3.3 Proposition for an asymmetric and sample-size sensitive entropy

3.3.1 How to estimate the probabilities

Instead of using classical frequency estimates, we carry out the estimates by mean of Laplace estimator which is given by $\lambda_i = \frac{Nf_i+1}{N+n}$

3.3.2 How to fix the "worst" distribution W

An important issue with asymmetric criterion is how can we determine the "most" uncertain reference distribution W? When the probability of each class is known, it is consistent to use these a priori probabilities of the classes. Otherwise, we could estimate them from the overall class frequencies in the learning dataset.

3.3.3 Asymmetric and sensitive entropy

Let $W = (w_1, w_2, \ldots, w_n)$ be the worst distribution, that has the maximal entropy value. The probabilities of the classes are estimated, locally, at each iteration of the growing process of the tree, by the Laplace estimator. The asymmetric entropy we propose is defined as follow :

Theorem

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

is an entropy measure that verifies the five properties cited above.

For the 2 classes problem, the Figure 2 shows the behavior of this function according to the parameters W and the size of the sample on which the probabilities are estimated.

Another non-centered entropy has been proposed in [14]. It results from a different approach that transforms the frequencies p_i's of the relevant node by means of a transformation that turns W into a uniform distribution. In the two class case, the transformation function is composed of two affine functions: $\pi = \frac{p}{2w}$ if $0 \leq p \leq w$ and $\pi = \frac{p+1-2w}{2(1-w)}$ if $w \leq p \leq 1$. The resulting non-centered entropy is then defined as the classical entropy of the transformed distribution. Though this method can be used with any kind of entropy measure, it is hardly extensible to more than two class problems.

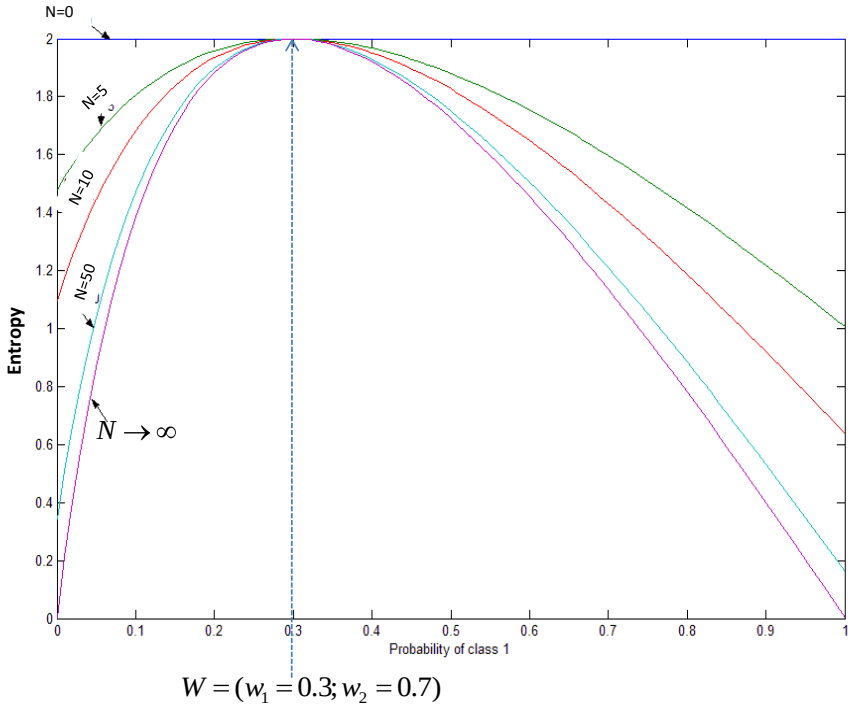

Fig. 2 Asymmetric and sample size sensitive entropy for 2 classes

4 Evaluation criteria of trees in the unbalanced case

4.1 Performance measures

There exist different measures for evaluating a prediction model. Most of them are based on the confusion matrix (see Table 1). Some measures are designed for the prediction of a specific modality (positive class) whereas the remaining modalities are gathered in the negative class : the recall rate ($\frac{TP}{TP+FN}$), that measures the rate of positive cases actually predicted as positive, and the precision rate ($\frac{TP}{TP+FP}$) that gives the proportion of real positive cases among those classified as positive by the classifier. The F-Measure is the harmonic mean of recall and precision. Other measures do not distinguish among outcome classes. We may cite here the overall error rate, and the sensibility and specificity (mean of recall and precision on each class). The latter measures are less interesting for us, since by construction they favor accuracy on the majority class. (Still, we may cite the PRAGMA measure [24] that allows the user to specify the importance granted for each class as well as its preferences in terms of recall and precision). It follows that recall and precision are the best suited measures when the concern is the prediction of a specific class, for instance rare class, most costly class, positive class and so on.

	Class +	Class -
Class +	True positives (TP)	False negatives (FN)
Class -	False positives (FP)	True negatives (TN)

Table 1 Confusion matrix for the two classes case.

The confusion matrix depicted in Table 1 is obtained for a decision tree by applying the relevant decision rule to each leaf. This is not a problem when the assigned class is the majority one. But with an asymmetric criterion this rule is not longer suited [15]: If we consider that the worst situation is a distribution W, meaning that the probability of class i is w_i in the most uncertain case, then no decision can be taken for leaves having this distribution. Hence, leaves where the class of interest is better represented than in this worst reference case ($f_i > w_i$) should be assigned to the class i. This simple and intuitive rule could be replaced by a statistical test, as we proposed it with the implication intensity [20] for instance. In this paper, we consider however the following simple decision rule: $C = i$ if $f_i > w_i$. This rule is adapted to the 2-class case. With k classes, the condition can indeed be satisfied for more than one modality and should then be reinforced. In [20] we proposed for instance to select the class with the lowest contribution to the off-centered entropy. To avoid the rule's limitation, we also move the decision threshold between 0 and 1 to observe the recall / precision graph. This allows us to see if a method dominates another one for different thresholds of decision, and can also help us to choose the most appropriate decision rule.

4.2 ROC curve

A ROC curve (Receiver operating characteristics) is a well suited tool for visualizing the performances of a classifier regarding results for a specific outcome class. Several works present its principles [7, 9]. First, a score is computed for each example. For decision trees, it is the probability to classify this example as positive. This probability is estimated by the proportion of positive examples in the leaf. Then, all examples are plotted in a false positive rate / true positive rate space, cumulatively from the best scored to the last scored. A ROC curve close to the main diagonal means that the model provides no useful additional information about the class. *A contrario* a ROC curve with a point in [0,1] means that the model perfectly separates positive and negative examples. The area under the ROC curve (AUC) summarizes the whole curve. We now examine how the ROC curve and the AUC may be affected when an asymmetric measure is used instead of a symmetric one.

4.3 Evaluations

4.3.1 Compared models and datasets

Our study is based on decision trees evaluated in 10 cross-validation to avoid the problems of over-fitting on the majority class. For each dataset we consider the quadratic entropy and the asymmetric entropy. The chosen stopping criterion, required to avoid over-fitting, is a minimal information gain of 3%. Other classical stopping criteria such as the minimal support of a leaf, or the maximal depth of the tree could be used. We selected the 11 datasets listed in Table 2. For each of them we have a two class outcome variable. We consider predicting the overall last frequent class. A first group of datasets is formed by strongly imbalanced datasets of the UCI repository [13]. In the dataset *letter* (recognition of hand-writing letters) we consider predicting the letter 'a' vs all the others (*letter_a*) and the vowels vs the consonants (*letter_vowels*). The classes of the dataset *Satimage* were merged into two classes as proposed by [5]. The datasets *Mammo1* and *Mammo2* are real data from the breast cancer screening and diagnosis collected within an industrial partnership. The goal is to predict from a set of predictive features whether some regions of interest on digital mammograms are cancers or not. This last example provides a good illustration of learning on a imbalanced dataset: Missing a cancer could lead to death, which renders the prediction of this class very important. A high precision is also requested since the cost of a false alarm is psychologically and monetary high.

Dataset	# of examples	# of features	Imbalance
Breast	699	9	34%
Letter_a	2000	16	4%
Letter_vowels	2000	16	23%
Pima	768	8	35%
Satimage	6435	36	10%
Segment_path	2310	19	14%
Waveform_merged	5000	40	34%
Sick	3772	29	6%
Hepatisis	155	19	21%
Mammo1	6329	1038	8%
Mammo2	3297	1038	15%

Table 2 Datasets.

4.3.2 Results and interpretation

Table 3 shows the AUC values obtained for each dataset. Figures 3,4,5,6 and 7 exhibit the ROC curves and the recall / precision graphs respectively for the datasets *Mammo1*, *Mammo2*, *Letter_a*, *Waveform_merged* and *Satimage*.

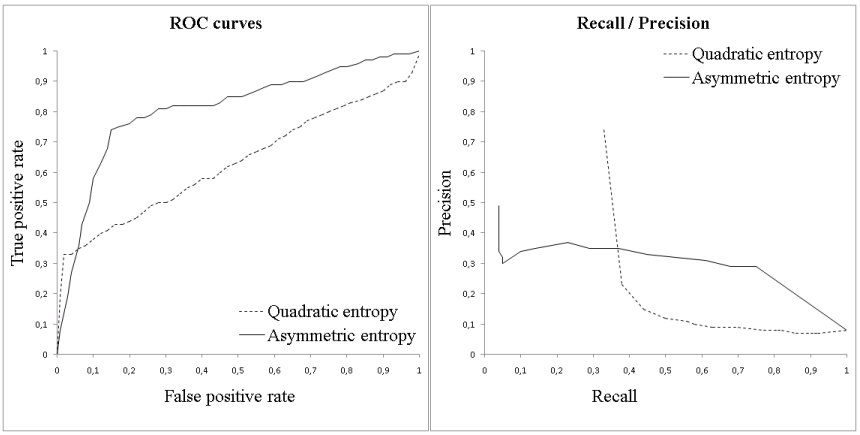

Fig. 3 Results for Mammo1

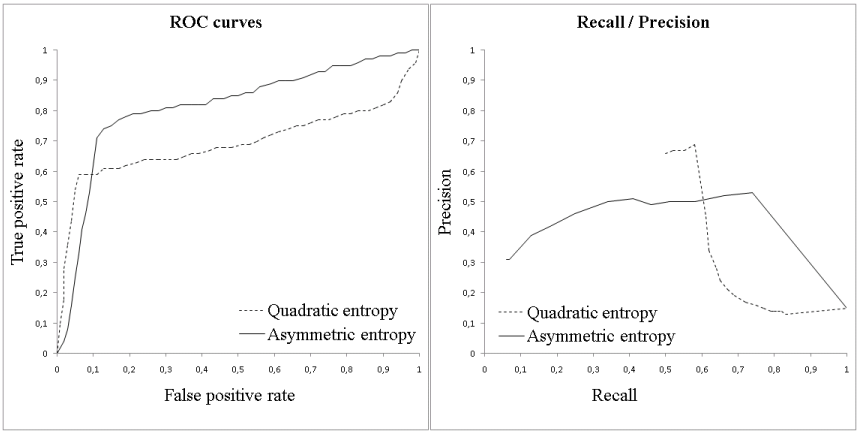

Fig. 4 Results for Mammo2

The recall / precision graphs show that when recall is high, the asymmetric criterion ends up with a better precision. This means that decision rules derived from a tree grown with an asymmetrical entropy are more accurate for predicting the rare class. On both real datasets (Figures 3 and 4) we see that if we try to maximize the recall (or to minimize the number of 'missed' cancers, or false negatives), we obtain fewer false positives with the asymmetric entropy. This is exactly the desired effect.

The ROC curve analysis shows that using the asymmetric entropy improves the AUC criterion (Table 3). More importantly, however is the form of the curves. The ROC curves of the quadratic entropy are globally higher on the left side of the graph, i.e. for high scores. Then the two ROC curves cross each other, and on the right side the asymmetric criterion is almost always dominating. We can thus conclude that

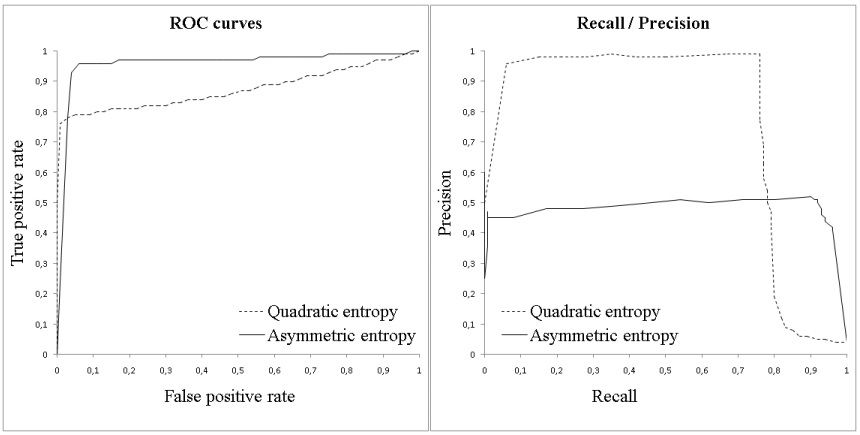

Fig. 5 Results for Letter_a

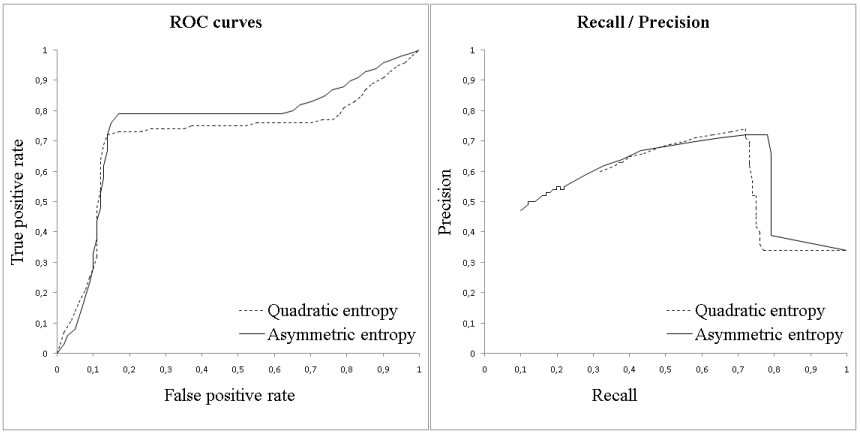

Fig. 6 Results for Waveform_merged

the lower the score, the more suited the use of an asymmetric entropy. As we have seen through several examples that when predicting rare events, we have to use small acceptation threshold (we accept a leaf when the observed frequency of the minority class exceeds the corresponding probability in the more uncertain distribution). Thus, ROC curves clearly highlight the usefulness of asymmetric entropies for predicting rare classes.

The two previous remarks mean that for seeking 'nuggets' of the minority class, we always get better recall and precision rates with an asymmetric criterion. In other words, if we accept predicting the class of interest with a score below 50%, then the smaller the score, the better the recall and precision rates when compared with those obtained with a symmetric criterion.

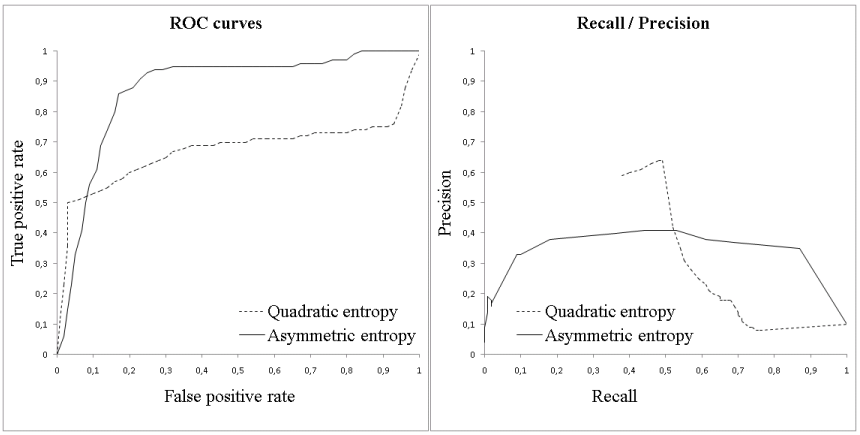

Fig. 7 Results for Satimage

Dataset	AUC with quadratic entropy	AUC with asymmetric entropy
Breast	0.9288	0.9359
Letter_a	0.8744	0.9576
letter_voyelles	0.8709	0.8818
pima	0.6315	0.6376
satimage	0.6715	0.8746
segment_path	0.9969	0.9985
Waveform_merged	0.713	0.749
sick	0.8965	0.9572
hepatisis	0.5554	0.6338
mammo1	0.6312	0.8103
mammo2	0.6927	0.8126

Table 3 Obtained AUC

5 Conclusion

We evaluated how using a splitting criterion based on an asymmetrical entropy to grow decision trees for imbalanced datasets influences the quality of the prediction of the rare class. If the proposed models are as expected less efficient in terms of global measures such as the error rate, ROC curves as well as the behavior of recall and precision as function of the acceptation threshold reveals that models based on asymmetric entropy outperform those built with a symmetric entropy, at least for low decision threshold.

For our empirical experimentation, the reference distribution W has been set up once and for all, as the a priori distribution of the probabilities estimated on the learning sample. A different approach would be to use at each node the distribution in the parent node as reference W. The criterion would in that case adapt itself at each node. A similar approach is to use Bayesian trees [4], where in each node we try to get rid of the parent node distribution. Finally, we noticed during our

experimentations that the choice of the stopping criterion is very important when we work on imbalanced datasets. Therefore, we plan to elaborate a stopping criterion suited for imbalanced data, that would, for instance, take into account the number of examples at each leaf, but allow for a lower threshold for leaves where the relevant class is better represented. In a more general way, various measures of the quality of association rules should help us to build decision trees.

We did not decide about the question of the decision rule to assign a class to each leaf. Since an intuitive rule is the one proposed in section 3, consisting in accepting the leaves where the class of interest is better represented than in the original distribution, we propose two alternative approaches: the first is to use statistical rules, or quality measures of association rules. The second is to use the graphs we proposed in this article, by searching optimal points on the recall / precision graph and on the ROC curve. We should consider the break-even Point (BEP, [21]) to find the best rate, or the Pragma criterion [24].

The extension of the concepts exposed in this article to the case of more than two modalities raises several problems. First, even if the asymmetric entropy applies to the multiclass case, some other measures are not. The problem of the decision rule is very complex with several classes. Indeed, setting a threshold on each class is not efficient, because this rule can be satisfied for several classes simultaneously. A solution is to choose the class with the frequency that departs the most from its associated threshold, or that with the smallest contribution to the entropy of the node. The methods of evaluation proposed in this paper (ROC curves and recall / precision graphs) are adapted for a class vs all the others, i.e. in the case with more than 2 classes, for the case where one modality among the others is the class of interest. It would be more difficult evaluating the model when two or more rare classes should be considered as equally relevant. The evaluation of multiclass asymmetric criteria will be the topic of future work.

References

1. Aczel, J., Daroczy, Z.: On Measures of Information and Their Characterizations. Academic Press, NY, S. Francisco, London (1975)
2. Barandela, R., Sanchez, J.S., Garcia, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(3) (2003) 849–851
3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification And Regression Trees. Chapman and Hall, New York (1984)
4. 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, ICDM04 (2004) 973978
5. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Technical Report 666, Berkeley, Department of Statistics, University of California (2004)
6. 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
7. Egan, J.: Signal detection theory and roc analysis. Series in Cognition and Perception (1975)

8. Elkan, C.: The foundations of cost-sensitive learning. In Nebel, B., ed.: IJCAI, Morgan Kaufmann (2001) 973–978
9. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letter 27(8) (2006) 861-874
10. Forte, B.: Why shannon's entropy. In Conv. Inform. Teor., 15 (1973) 137–152
11. Hartley, R.V.: Transmission of information. Bell System Tech. J. 7 (1928) 535–563
12. Hencin, A.J.: The concept of entropy in the theory of probability. Math. Found. of Information Theory (1957) 1–28
13. Hettich, S., Bay, S.D.: The uci kdd archive (1999)
14. Lallich, S., Lenca, P., Vaillant, B.: Probabilistic framework towards the parametrisation of association rule interestingness measures. Methodology and Computing in Applied Probability 9(3) (2007) 447–463
15. Marcellin, S., Zighed, D., Ritschard, G.: An asymmetric entropy measure for decision trees. 11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France (2006) 1292–1299
16. Provost, F.: Learning with imbalanced data sets. Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets (2000)
17. 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
18. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
19. Renyi, A.: On measures of entropy and information. 4th Berkely Symp. Math. Statist. Probability 1 (1960) 547–561
20. Ritschard, G., Zighed, D., Marcellin, S.: Donnes dsquilibres, entropie dcentre et indice d'implication. In Gras, R., Orus, P., Pinaud, B., Gregori, P., eds.: Nouveaux apports thoriques l'analyse statistique implicative et applications (actes des 4mes rencontres ASI4, 18-21 octobre 2007), Castellon de la Plana (Espana), Departament de Matematiques, Universitat Jaume I (2007) 315–327
21. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1) (2002) 1–47
22. Shannon, C.E.: A mathematical theory of communication. Bell System Tech. J. 27 (1948) 379–423
23. Shannon, C.A., Weaver, W.: The mathematical of communication. University of Illinois Press (1949)
24. Thomas, J.: Apprentissage supervis de donnes dsquilibres par fort alatoire. Thse de doctorat, Universit Lyon 2 (2009)
25. Zighed, D.A., Marcellin, S., Ritschard, G.: Mesure d'entropie asymtrique et consistante. In Noirhomme-Fraiture, M., Venturini, G., eds.: EGC. Volume RNTI-E-9 of Revue des Nouvelles Technologies de l'Information., Cpadus-Editions (2007) 81–86
26. Zighed, D., Rakotomalala, R.: Graphe d'induction: Apprentissage et Data Mining. Herms, Paris (2000)