An asymmetric entropy measure for decision trees

Simon Marcellin
Laboratoire ERIC
Université Lumière Lyon 2
5 av. Pierre Mendès-France
69676 BRON Cedex FRANCE
simon.marcellin@univ-lyon2.fr

Djamel A. Zighed
Laboratoire ERIC
Université Lumière Lyon 2
5 av. Pierre Mendès-France
69676 BRON Cedex FRANCE
zighed@univ-lyon2.fr

Gilbert Ritschard
Department of Econometrics,
University of Geneva,
40 bd du Pont-d'Arve,
CH-1211 Geneva 4, Switzerland
gilbert.ritschard@themes.unige.ch

Abstract

In this paper we present a new entropy measure to grow decision trees. This measure has the characteristic to be asymmetric, allowing the user to grow trees which better correspond to his expectation in terms of recall and precision on each class. Then we propose decision rules adapted to such trees. Experiments have been realized on real medical data from breast cancer screening units.

Keywords: Decision trees, entropy, class imbalance, asymmetric learning, decision rules.

1 Introduction

During standard decision trees growing, we may notice two properties linked one to each other: on the one hand, the symmetry of the splitting criterion, for instance the Shannon entropy [12] or the Gini measure [4], which implies that maximal uncertainty is reached for equiprobability distribution over the classes; on the other hand, the application of the majority rule for assigning a leaf of the tree to a specific class. In many problems, prior distributions classes are not balanced and may not have the same importance [8]. This happens for example in marketing [5], medical computer-aided diagnosis or fraud detection [1], where some classes are much more important than others. In medical fields, missing a cancer (false negative) doesn't have the same consequences than predicting wrong cancer (false positive). Thus users' requirements towards a classification model are different. Let us explain this issue in more details. We consider an example of breast cancer computer-aided diagnosis where the aim is to label regions on digitized films as "cancer" or "non-cancer". In this framework the "cancer" class is much less represented in the datasets, but it is imperative to classify all them well. A standard decision tree considers that maximum uncertainty is reached in leaves having 50% of "cancer" and 50% of "non-cancer". Likewise, the leaves are classified as "cancer" as soon as "cancer" proportion exceeds 50% and "non-cancer" otherwise. We can see in this example that this is not suitable to this kind of problems. Indeed as a "non-cancer" decision may have serious sequel, we could decide to assign to this category only the leaves with a "non-cancer" proportion greater than 90%, in order to avoid false positive results. On the other hand, it is so important to find all cancers that we could decide to classify leaves as "cancer" from a proportion of say 20%. If we fix these two proportions as decision thresholds, we give rise to an indecision area for "cancer" proportion p between 10% and 20%. Some approaches have proposed to deal with this problem. We can class them into four categories [1]. First, the cost-sensitive approaches allow penalizing some types of errors, balancing the number of examples of the concerned class. Then, sampling methods allow over-representing the minority class or under-representing the majority class [11, 14]. Some wrapper methods as MetaCost have also been proposed [7]. They produce several instances of a classifier through bootstrap, re-label each example by votes and build another model using the new labels,. Finally, the methods closer to the one proposed in this paper try to include a bias directly in the splitting criterion, in particular by using a cost function instead of a classical entropy criterion [9, 6]. Moreover, the majority decision rules for labelling the leaves must be modified according the two thresholds mentioned before. This does not change the produced tree but just the decision for each leaf. In order to have an entirely coherent model, the splitting criterion must be suitable to manage the fact that the maximum uncertainty situation corresponds to the indecision situation in the decision rules. For example, in our two-class problem, if we class a leaf as "cancer" when the frequency of cancers exceeds 20% and as "non-cancer" when their frequency exceeds 90%, the maximal uncertainty must be reached when the proportion of "cancer" class is between 10% and 20%.

We will list, in section 2, the properties requested for a new criterion. Section 3 presents our asymmetric entropy criterion. In section 4 we present how to adapt the decision rules to this asymmetric entropy measure. Section 5 details the results achieved with a real medical dataset within the framework of a computer-aided diagnosis of breast cancer system, and two standard datasets from the UCI repository [9]. Finally, section 6 concludes and proposes some extensions to our work.

2 Expected properties of an asymmetrical uncertainty measure in the two class case

Let *p* design the probability to be a "cancer", 1-*p* being for the probability of "non cancer". The usual splitting criteria that have been proposed are symmetrical, implying that maximal uncertainty is reached for the equiprobability distribution [15], i.e. for *p*=0.5 in the two class case.

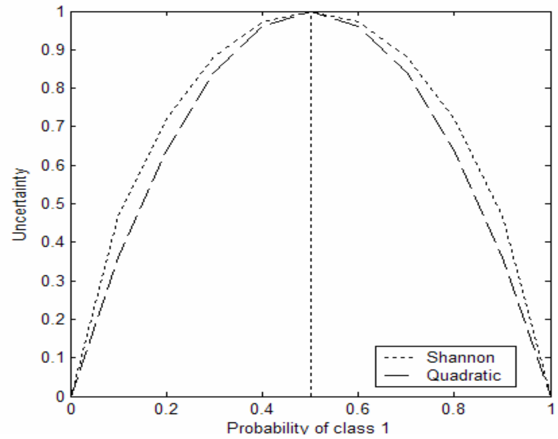

Figure 1: Examples of standard entropy measures for a two-class problem

Yet, we need the maximal uncertainty to be reached for a given probability of "cancer", noted *p*=*w*. This criterion should verify the classical properties of the entropy measures shown bellow. More formally, we seek a non negative function *h* of *p*. In real application, *p* is estimated at each leaf by the frequency. This function *h* should respect the entropy properties, except that the maximum should be reached for *p*=*w* instead of 0.5. So the requested properties are [15]:

1. Strict concavity

$$(1)\ \frac{\partial^2 h}{\partial p^2} < 0$$

2. Minimality:

$$(2)\ h(p=1)=0 \ \ \&\ \ h(p=0)=0$$

3. Maximality

$$(3)\ \frac{\partial h}{\partial p} = 0\ \ ;\ \ \text{for}\ \ p = w$$

We assume there exists a rational function verifying the three previous conditions which could be expressed as follow :

$$(4)\ h(p) = \frac{ap^2 + bp + c}{dp + e}$$

Where *a*,*b*,*c*,*d* and *e* are the coefficients to be found. To remove a degree of freedom, we consider also the following additional constraint on the maximum value:

(5) $h(p=w)=1$

Using constraints (2) and (5), the function (4) simplifies to:

$$h(p)=\frac{-p^{2}+p}{(-2w+1)p+w^{2}}=\frac{p(1-p)}{(-2w+1)p+w^{2}}$$

where the reference probability *w* is given by the user. With this function we may represent the contribution of a class to the global uncertainty of a node. It verifies the requested properties.

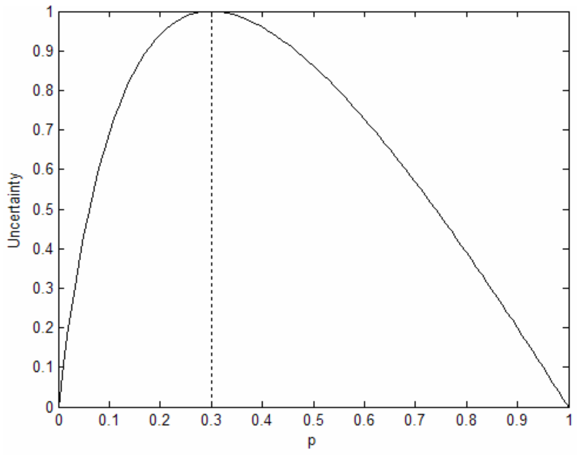

Figure 2: Asymmetric uncertainty measure for *w*= 0.3

3 Asymmetric entropy

So we have determined an uncertainty function for two classes. One way of extending this function to the *k* class case is by considering the additively separable function:

$$H(p_1,p_2,\ldots,p_i,\ldots,p_k)=\sum_{i=1}^{k}h(p_i)$$

Where

$$h(p_i)=\frac{p_i(1-p_i)}{(-2w_i+1)p_i+w_i^{2}}$$

By construction, *H* reaches is maximum at

$\mathbf{w} = (w_1,\ldots,w_k)$.

Figure 3 shows the value of the entropy for a three-class problem, in function of p_1 and p_2 (the third class probability is implicit because $p_3=1-p_1-p_2$).

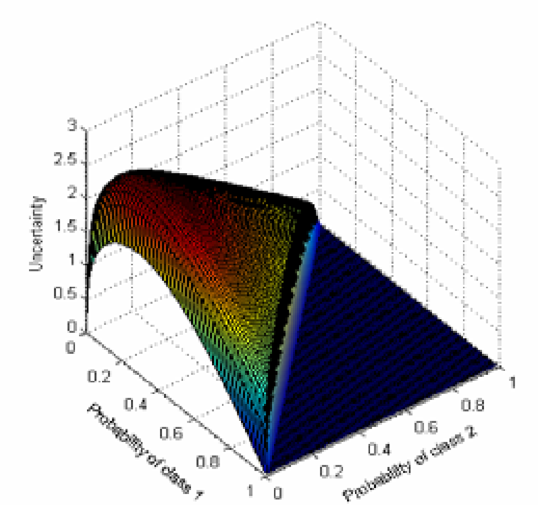

Figure 3: Three-class asymmetric entropy,

We can now grow a decision tree taking into account users' requirements. At each new split in the decision tree, we try to get rid of the uncertainty situations represented by our asymmetric entropy measure. The aim of the asymmetric entropy is to produce a tree tailored as much as possible to the priors **w** provided by the user.

4 Impact on decision rules

How can we classify a leaf from its distribution $P=\begin{bmatrix}p_1 & \ldots & p_k\end{bmatrix}$? Standard trees set the class *C* with the majority rule:

(R0) $C=i$ if $p_i>p_j\forall j\neq i$

According to the construction of our entropy measure, we propose alternative approach in order to take into account the fact that there is an area delimited by some thresholds in which the decision is uncertain.

To fix the ideas without loosing generality, let us consider that two thresholds:

δ_1 and δ_2 ; $\delta_i\in[0,1]$ have been fixed by the user. The threshold δ_2 is the minimal proportion at which we conclude for the class "cancer" and the threshold δ_1, the proportion bellow which we conclude for the complementary class "non-cancer". The decision rule for labelling a leaf reads then as follows:

- If $p>\delta_1$ then the class is "cancer"
- If $p<\delta_2$ then the class is "non-cancer"

This rule leads us to an interval δ=$\left[\delta_1, \delta_2\right]$ in which we cannot conclude for any class. Figure 3 shows the situation.

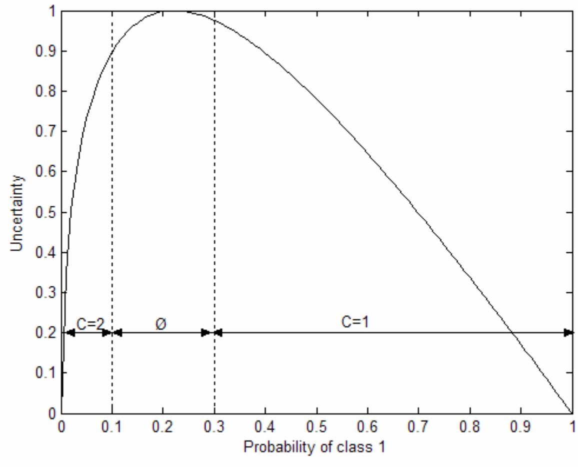

Figure 4: Decision rules for a two-class problem, δ=[0.1, 0.3]

5 Experiments

We have tested our method on real data from breast cancer screening units. The aim is to detect tumours on digitized mammograms. Several hundreds of films have been annotated by radiologists. Then, those films have been segmented with imaging methods, in order to obtain regions of interest. Those zones have been labelled as "cancer" if they correspond to a radiologist annotation or "non-cancer" otherwise. At last, a large number of features (based on grey-level histogram, shape and texture) have been computed. The aim is to build a model able to separate cancers and non-cancers. (See the description of the mammo dataset table 1).

Table 1: Description of the mammo dataset

Dataset	No. of features	No. of examples	% of cancer
Mammo	970	850	2%

To allow comparisons with other approaches, we have also tested our method on two standard machine learning datasets [9, 6]. On these imbalanced datasets we tried to get the best recall on the minority class, keeping a correct precision on this class.

Table 2: Description of the UCI datasets

Dataset	No. of features	No. of examples	% of the small class
Hypothyroid	28	3772	8%
Satimage	36	6435	10%

With each dataset we have built three series of classifiers using 10 folds cross-validation: standard C4.5 [13] and Random Forest [2, 3]; C4.5 and Random Forest using cost matrix with the WEKA software [14]; and tree and Random Forest with the asymmetric entropy criterion we proposed. We also compare our results with those obtained by [6]. For each test we present recall and precision on the minority class. As we present only two-class problems, the results on the majority class are implicit. That's why we do not give them here.

Table 3: Results for mammo

Methods	Recall	Precision
C4.5	0.18	0.23
Random Forest	0.0	0.0
C4.5 with cost	0.5	0.08
C4.5 with cost*	0.11	0.25
RF with cost	0.67	0.03
RF with cost*	0.61	0.42
Asymmetric tree	0.11	0.17
Asymmetric RF	0.88	1.0

For asymmetric methods and cost matrix we used several sets of parameters. Tables 4, 5 and 6 present the best results obtained for each method. We may see that the best results are always obtained with an asymmetric Random Forest. For our mammo dataset, asymmetric entropy with a single decision tree does not improve results (cost-sensitive C4.5 is better), but causes real enhancement with the Random Forest.

Table 4: Results for hypothyroid

* Using an alternative set of parameters

Methods	Recall	Precision
C4.5	0.98	0.98
Random Forest	0.96	0.96
C4.5 with cost	0.98	0.94
C4.5 with cost*	0.98	0.96
RF with cost	0.99	0.93
RF with cost*	0.99	0.96
Asymmetric tree	0.98	0.97
Asymmetric RF	1.0	0.98
BRF	0.95	0.63
WRF	0.93	0.83

For the hypothyroid dataset, we also present the results got with Balanced Random Forest (BRF) and Weighted Random Forest (WRF) [6]. On this dataset, asymmetric Random Forest dominates all other methods, whatever the parameters.

Table 5: Results for satimage

Methods	Recall	Precision
C4.5	0.55	0.59
Random Forest	0.54	0.84
C4.5 with cost	0.92	0.29
C4.5 with cost*	0.6	0.52
RF with cost	0.95	0.33
RF with cost*	0.64	0.73
Asymmetric tree	0.93	0.3
Asymmetric tree*	0.71	0.49
Asymmetric RF	0.99	0.29
Asymmetric RF*	0.91	0.96
BRF	0.67	0.64
BRF*	0.77	0.56
WRF	0.69	0.69
WRF*	0.77	0.61

With the satimage dataset, we have retained the minority class and merged the five others together to create the majority class. We also present BRF and WRF results.

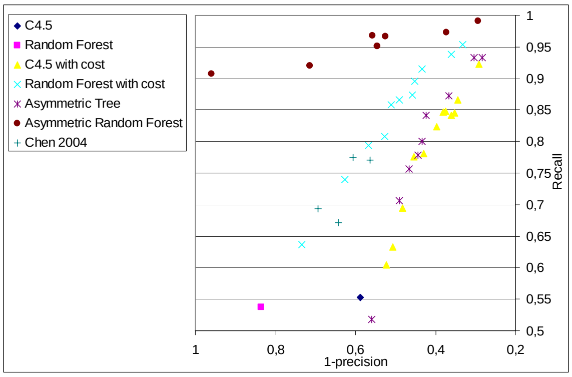

Figure 5: Recall and precision for satimage with different parameters

We observe on figure 5 that for the satimage dataset, the asymmetric Random Forest dominates other methods again. The Random Forest with cost matrix is very close to WRF and BRF. Finally we may notice that for this dataset, asymmetric tree is a bit better than C4.5 using cost matrix.

6 Conclusion and future works

We propose an asymmetric entropy measure for decision trees. By using it with adapted decision rules, we can grow trees taking into account the user's specifications in terms of recall and precision for each class, and thus obtain better results on imbalanced datasets. We can notice that this method entails no additional computing complexity, and that the parameters can be set intelligibly by the user. In the future we plan to improve our works on different points. Particularly, we will conduct a theoretical comparison between this approach and the ones that try to introduce costs in the splitting criterion. Indeed, as those two approaches can be expressed in the same way, it seems to us that using costs entails ruptures in the concavity of the splitting criterion, which could produce under-optimal trees. Furthermore, the entropy that we propose can be enhanced, in particular the aggregation of the uncertainty (we use a sum) for each class. Finally, experiments with more than two class problems will be led.

Acknowledgements

This work has been realized within a thesis cofinanced by the French Ministry of Research and Industry. We would like to thank the ARDOC's breast cancer screening units and the senology unit of Centre-République (Clermont-Ferrand, France) for their expertise and the mammography data. We also thank Pierre-Emmanuel Jouve, Julien Thomas and Jérémy Clech (Fenics company, Lyon, France), for their help and advices that allowed us to realize this research.

References

[1] R. Barandela, J. S. Sanchez, et al. (2003). Strategies for learning in class imbalance problems, In *Pattern Recognition*, volume 36(3), pages 849-851.

[2] L. Breiman (2001). Random Forests, In *Machine Learning*, volume 45(1), pages 5-32.

[3] L. Breiman (2002). Looking inside the black box. In *WALD lectures, the 277th meeting of the Institute of Mathematical Statistics*, Banff, Alberta, Canada.

[4] L. Breiman, J. H. Friedman, et al. (1984). Classification and Regression Trees. Belmont, Wadsworth.

[5] J.-H. Chauchat, R. Rakotomalala, et al. (2001). Targeting Customer Groups using Gain and Cost Matrix: a Marketing Application. In *Proceedings of the "Datamining and Marketing" Whorkshop, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, pages 1-14.

[6] C. Chen, A. Liaw, et al. (2004). Using Random Forest to Learn Imbalanced Data. Technical Report. Berkeley, Department of Statistics, University of California.

[7] P. Domingos (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 155-164.

[8] C. Elkan (2001). The Foundations of Cost-Sensitive Learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973-978.

[9] S. Hettich and S. D. Bay (1999). The UCI KDD Archive. Irvine, california, USA, University of California, Department of Information and Computer Science.

[10] C. X. Ling, Q. Yang, et al. (2004). Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, 69, Banff, Alberta, Canada.

[11] F. Provost (2000). Learning with Imbalanced Data Sets. In *Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets*.

[12] J. R. Quinlan (1986). Induction of Decision Trees, In *Machine Learning*, volume 1(1), pages 81-106.

[13] J. R. Quinlan (1993). C4.5: programs for machine learning. San Francisco, Morgan Kaufmann Publishers Inc.

[14] I. H. Witten and E. Frank (2005). Data Mining: Practical machine learning tools and techniques. San Francisco.

[15] D. Zighed and R. Rakotomalala (2000). Graphe d'induction Apprentissage et Data Mining. Paris, Hermès.