Detection of breast cancer using an asymmetric entropy measure

Simon Marcellin[1], Djamel A. Zighed[1], and Gilbert Ritschard[2]

[1] Laboratoire ERIC, Université Lumière Lyon 2, 5 av. Pierre Mendès-France 69676 Bron Cedex France {abdelkader.zighed,simon.marcellin}@univ-lyon2.fr
[2] Department of Econometrics, University of Geneva, 40 bd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland gilbert.ritschard@themes.unige.ch

Summary. 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.

Key words: entropy measures, decision trees, classification, asymmetric error

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 [Qui86] or the Gini measure [BFO84], 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 [Elk01]. This happens for example in marketing [CRC01], medical computer-aided diagnosis or fraud detection [BSG03], 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. Some approaches have proposed to deal with this problem. We can class them into four categories [BSG03]. 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 [Pro00, WF05]. Some wrapper methods as MetaCost have also been proposed [Dom99]. 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 [HB99, CLB04]. 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 10% and as ”non-cancer” when their frequency exceeds 90%, the maximal uncertainty must be reached when the proportion of ”cancer” class is 10%.

We present, in section 2, our asymmetric entropy criterion and its properties. In section 3 we present how to adapt the decision rules to this asymmetric entropy measure. Section 4 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 [HB99]. Finally, section 5 concludes and proposes some extensions to our work.

2 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 [ZR00], i.e. for $p = 0.5$ in the two class case (see figure 1).

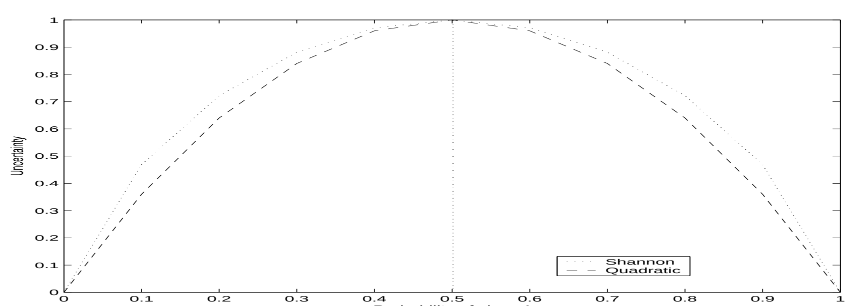

Fig. 1. Examples of standard entropy measures for a two-class problem

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 and continuous function of p depending on the parameter w, noted $h_w(p)$. In real application, p is estimated at each leaf by the frequency. This function h_w should respect the entropy properties, except that the maximum should be reached for $p = w$ instead of 0.5. So the requested properties are [ZR00]:

1. Strict concavity

$$\frac{\partial^2 h_w}{\partial p^2} < 0 \tag{1}$$

2. Minimality

$$h_w(p = 0) = 0 \ \& \ h_w(p = 1) = 1 \tag{2}$$

3. Maximality

$$\frac{\partial h_w}{\partial p} = 0; \text{ for } p = w \tag{3}$$

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

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

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:

$$h_w(p = w) = 1 \tag{5}$$

Using the constraints 2, 3, 1 and 5, the function 4 simplifies to:

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

Where the reference probability w is given by the user. With this function we may represent the uncertainty for a given probability distribution in the two class case. It verifies the requested properties. At each step of the decision tree growth, we will use this criterion to evaluate the different features.

3 Impact on decision rules

How can we classify a leaf from its distribution $P = [p_1, p_2]$? Standard trees set the class C with the majority rule:

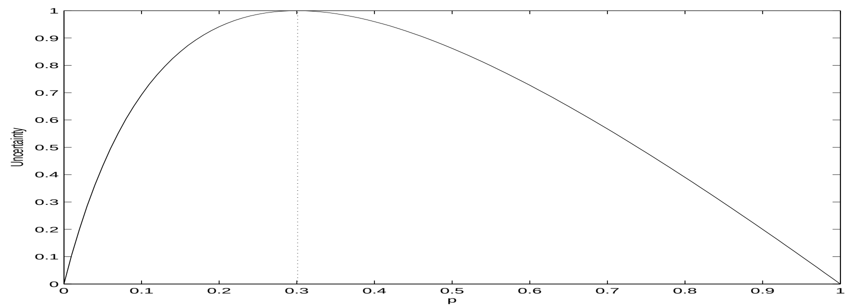

Fig. 2. Asymmetric uncertainty measure for w = 0.3

$$C = i \text{ if } p_i > p_j \; \forall \, j \neq i \tag{6}$$

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 ($\delta_i \in [0, 1]$, $\delta_1 \leq \delta_2$) 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 $\delta = [\delta_1, \delta_2]$ in which we cannot conclude for any class. if $\delta_1 = \delta_2$, one obtain an unique indecision point for $p = w$. Figure 3 shows the situation.

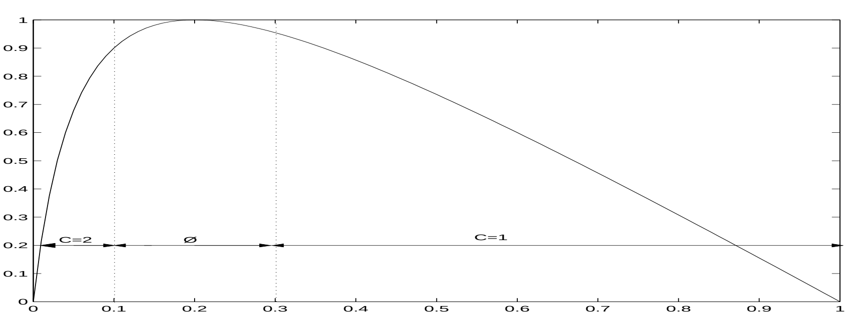

Fig. 3. Decision rules for a two-class problem, $w = 0.2$, $\delta = [0.1, 0.3]$

The parameters of our methods, i.e. δ_1 and δ_2 as well as w, should be set in function of the dataset imbalance on the one hand and the user's requirement in term of recall and precision on the other.

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

Table 1. Description of the datasets

Dataset	Features	Examples	Proportion of the minority class
Mammo	115	3478	10%
Hypothyroid	28	3772	8%
Satimage	36	6435	10%

With each dataset we have tested C4.5 [Qui93], J48 with cost matrix proposed by the WEKA software [WF05], and trees with the asymmetric entropy criterion we proposed. For each test we present recall and precision rates 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. It is also important to clarify that for all our experiments, we use $\delta_1 = \delta_2 = w$, i.e. we don’t define any indecision area.

Table 2. Results for mammo

Methods	Parameters	Recall	Precision
C4.5	-	0.02	0.42
C4.5 with cost	cost 1:2	0.51	0.45
C4.5 with cost	cost 1:10	0.59	0.36
Asymmetric tree	$w = 0.6$	0.61	0.61
Asymmetric tree	$w = 0.97$	0.59	0.33

For asymmetric methods and cost matrix we used several sets of parameters. Tables 2, 3 and 4 present the best results obtained for each method.

With the satimage dataset, we have retained the minority class and merged the five others together to create the majority class.

We notice that our method allows us to get a better recall rate on the minority class, with a loss of precision. The results are close to those obtained with cost matrices, but we may see three advantages of our method:

Table 3. Results for hypothyroid

Methods	Parameters	Recall	Precision
C4.5	-	0.98	0.98
C4.5 with cost	cost 1:10	0.98	0.95
C4.5 with cost	cost 1:30	0.98	0.91
Asymmetric tree	$w = 0.8$	0.98	0.95
Asymmetric tree	$w = 0.9$	0.99	0.79

Table 4. Results for satimage

Methods	Parameters	Recall	Precision
C4.5	-	0.55	0.59
C4.5 with cost	cost 1:5	0.6	0.52
C4.5 with cost	cost 1:10	0.63	0.51
C4.5 with cost	cost 1:80	0.85	0.35
Asymmetric tree	$w = 0.55$	0.46	0.64
Asymmetric tree	$w = 0.7$	0.64	0.53
Asymmetric tree	$w = 0.95$	0.92	0.3

- For the same recall rate, our method gives a bit better precision rate.
- By modifying the parameter w, we are able to get very high recall or precision rates. This allows the user to build models adapted to his goals.
- Contrary to the costs, the parameter w is understandable: it is the worst distribution for a given problem.

These are preliminary tests which should be detailed in the future. The next section presents extensions of this work.

5 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. We may 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 consider the cases with more than two classes. Indeed, this measure may be adapted to such cases using a sum aggregation. The user will just have to set the "worst" prior distribution $W = [w_1, ..., w_k]$ where k is the number of classes, so this should be easier than the cost matrix when the number of classes is high. We will also detail the experiments in the two-class case, and try to get better results by using a more adapted stopping criterion in the decision tree growing. We will also 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. At last, we will experiment Random Forest [Bre01,Bre02] with our measure to compare our method with those proposed by Chen and Liu [CLB04].

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

[BSG03] Barandela, R., Sanchez, J.S., Garcia, V. and Rangel, E.: Strategies for learning in class imbalance problems, In Pattern Recognition, 36(3), 849-851 (2003)

[Bre01] Breiman, L.: Random Forests, In Machine Learning, 45(1), 5-32 (2001)

[Bre02] Breiman, L.: Looking inside the black box. In WALD lectures, the 277th meeting of the Institute of Mathematical Statistics, Banff, Alberta, Canada (2002)

[BFO84] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone., C.J.: Classification and Regression Trees. Belmont, Wadsworth (1984)

[CRC01] Chauchat, J.-H., Rakotomalala, R., Carloz, M. and Pelletier, C.: 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), 1-14 (2001)

[CLB04] Chen, C., Liaw, A. and Breiman, L.: Using Random Forest to Learn Imbalanced Data. Technical Report. Berkeley, Department of Statistics, University of California (2004)

[Dom99] Domingos, P.: MetaCost: A general method for making classifiers cost-sensitive. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), 155-164 (1999)

[Elk01] Elkan, C.: The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01), 973-978 (2001)

[HB99] Hettich, S. and Bay, S.D.: The UCI KDD Archive. Irvine, california, USA, University of California, Department of Information and Computer Science. (1999)

[Pro00] Provost, F.: Learning with Imbalanced Data Sets. In Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets (2000)

[Qui86] Quinlan, J.R.: Induction of Decision Trees, In Machine Learning, 1(1), 81-106 (1986)

[Qui93] Quinlan, J.R.: C4.5: programs for machine learning. San Francisco, Morgan Kaufmann Publishers Inc. (1993)

[WF05] Witten, I.H. and Frank, E.: Data Mining: Practical machine learning tools and techniques. San Francisco (2005)
[ZR00] Zighed, D. and Rakotomalala, R.: Graphe d'induction Apprentissage et Data Mining, Hermès (2000)