Evaluating decision trees grown with asymmetric entropies

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

[1] Université Lumière Lyon 2 Laboratoire ERIC, Bât L, Campus Porte des Alpes
5, av. Pierre Mendès-France, F-69600 Bron, France
{abdelkader.zighed,simon.marcellin}@univ-lyon2.fr http://eric.univ-lyon2.fr
[2] Université de Genève, Département d'économétrie, 40 bd du Pont-d'Arve
CH-1211 Geneva 4, Switzerland, gilbert.ritschard@unige.ch

Abstract. We propose to evaluate the quality of decision trees grown on imbalanced datasets with a splitting criterion based on an asymmetric entropy measure. To deal with the class imbalance problem in machine learning, especially with decision trees, different authors proposed such asymmetric splitting criteria. After the tree is grown a decision rule has to be assigned to each leaf. The classical Bayesian rule that selects the more frequent class is irrelevant when the dataset is strongly imbalanced. A best suited assignment rule taking asymmetry into account must be adopted. But how can we then evaluate the resulting prediction model? Indeed the usual error rate is irrelevant when the classes are strongly imbalanced. Appropriate evaluation measures are required in such cases. We consider ROC curves and recall/precision graphs for evaluating the performance of decision trees grown from imbalanced datasets. These evaluation criteria are used for comparing trees obtained with an asymmetric splitting criterion with those grown with a symmetric one. In this paper we only consider the cases involving 2 classes.

1 Introduction

Learning from imbalanced datasets is an important issue in datamining [1, 2]. A dataset is imbalanced when the distribution of the modalities of the class variable is far away from the uniform distribution. 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 breakdown; or in the bank field, to predict insolvent costumers or frauds in transactions. In these cases, there is one rare state of the class variable (ill, breakdown, insolvent, fraud) that should be detected in priority. 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. that which minimizes the global error rate. 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.

First, 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 or quadratic entropy. 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 that for which we have the same number of examples 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 1% of the people are sick, ending with a leaf in which 50% of the members are sick would be very interesting and would carry a lot of information for the user of the model. 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.

In decision trees, the imbalance of the prediction class influences the learning process during these two steps. This paper focuses on the first issue. Asymmetric criterion were proposed to deal with this imbalance aspect in decision trees. How do such criteria influence the learning? If we use an asymmetric criterion, what performance measure should be used to evaluate the gain of using this criterion? Our proposition is to consider ROC curves and recall/precision graphs and apply them for evaluating the gain brought by using splitting criteria based on asymmetric measures over those based on symmetrical measures. In section 2 we present the decision trees and the asymmetric criterion. In section 3 we propose evaluation methods to compare trees built with a symmetric criterion versus those grown with an asymmetric one. Section 4 presents the methodology of our evaluation and exposes our results. We finish by the section 5 that concludes and proposes future works.

2 Asymmetric criteria for decision trees

2.1 Notations and basic concepts

We note 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 $\mathcal{C}$. Letting m_j be the number of different values taken by X_j and n the number of modalities of C, we have $\mathcal{C} = \{c_1, \ldots, c_n\}$. And when it is not ambiguous, we

denote the class c_i simply by i. Algorithms of trees induction generate a model $\phi(X_1, \ldots, X_p)$ for the prediction of C represented by a decision tree [3, 4] or an induction graph [5]. 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 know only 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. The notion of entropy is defined mathematically by axioms out of the context of machine learning. See for instance [6] and [7] for details. The entropy H on the partition S to minimize is generally a mean entropy such as $H(S) = \sum_{s \in S} p(s) h(p(1|s), \ldots, p(i|s), \ldots, p(n|s))$ 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 $H_n(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$ .. By continuity we set $0 \log_2 0 = 0$. There are other entropy measures [8] 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.

2.2 Asymmetric criteria

The properties of classical entropy measures such as those cited above (Shannon, quadratic) are not suited to inductive learning for reasons exposed in [5]. First, the uniform distribution is not necessarily the most uncertain. Second, the computation of the entropy being based on estimates of the probabilities, it should account for the precision of those estimates, i.e. account for the sample size. That is why we proposed in [5] a new axiomatic leading to a new family of more general measures allowing for a user defined maximal entropy reference and sensitive to the sample size. Let $\lambda_i = \frac{Nf_i+1}{N+n}$ be the Laplace estimator of p_i, $W = (w_1, w_2, \ldots, w_n)$ the vector with maximal entropy and N the sample size. The asymmetric entropy we proposed reads:

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

An other non-centered entropy has been proposed in [9]. 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 the more than two class problem.

3 Evaluation criteria of trees in the imbalanced case

3.1 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: the recall rate ($\frac{TP}{TP+FN}$)and the precision rate ($\frac{TP}{TP+FP}$) The F-Measure is the harmonic mean of recall and precision. Other measures do not distinguish among outcome classes. We may cite here 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 [10] 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 (rare) class as in our setting.

	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 [11]: 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, has we proposed it with the implication intensity [12] 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. 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 an other one for different thresholds of decision, and can also help us to choose the most appropriate decision rule.

3.2 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 [13, 14]. 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 separates perfectly 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 Evaluations

4.1 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, would preterite the minority class. 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 [15]. 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 Chen and Liu [16]. 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.

4.2 Results and interpretation

Table 3 shows the AUC values obtained for each dataset. Figures 1,2,3,4 and 5 exhibit the ROC curves and the recall / precision graphs respectively for the datasets *Mammo1*, *Mammo2*, *Letter_a*, *Waveform_merged* and *Satimage*.

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

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.

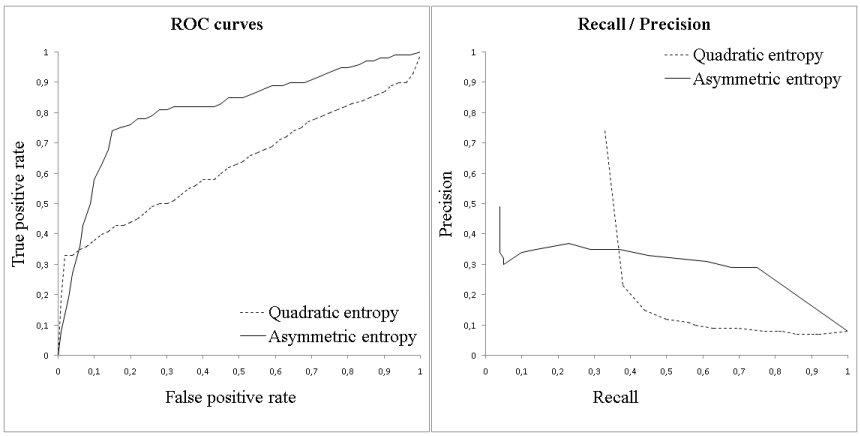

Fig. 1. Results for Mammo1

the rare class. On both real datasets (Figures 1 and 2) 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 curves analysis shows that using the asymmetric entropy improves the AUC criterion (Table 3). More importantly is however 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 the lower the score, the more suited the use of an asymmetric entropy. We saw in section 2 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

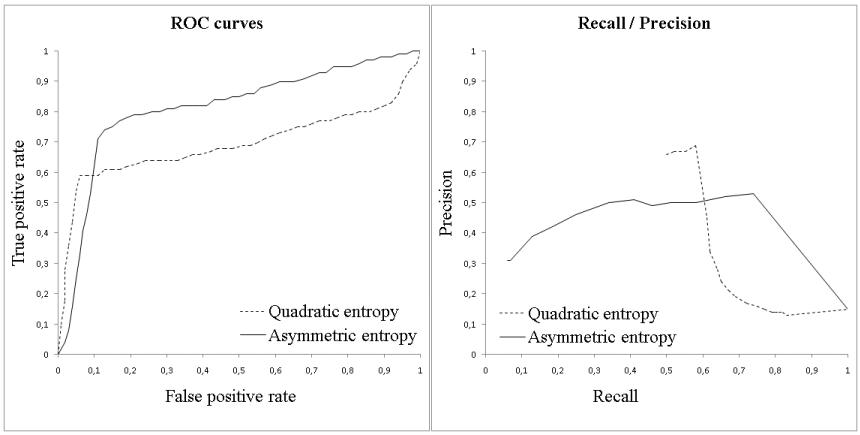

Fig. 2. Results for Mammo2

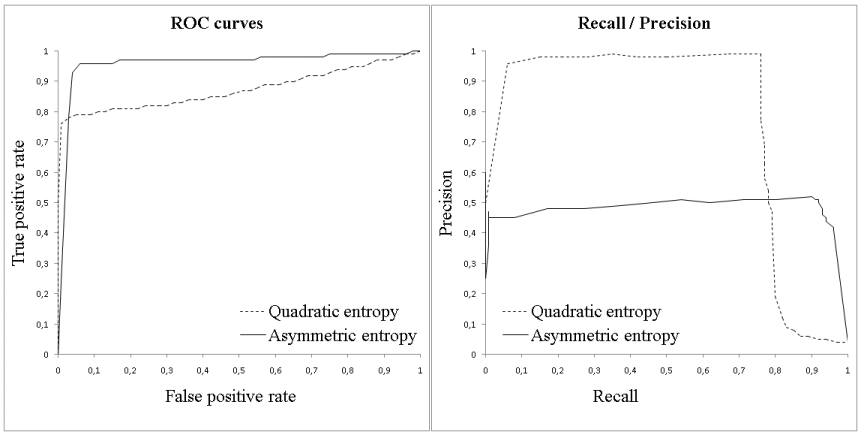

Fig. 3. Results for Letter_a

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.

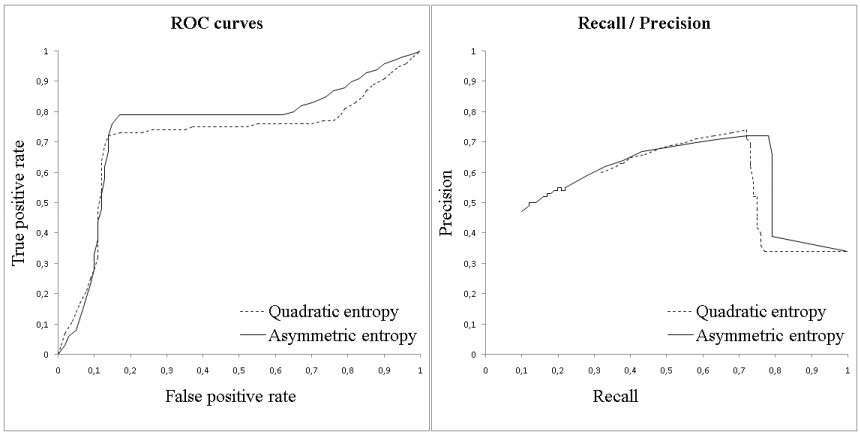

Fig. 4. Results for Waveform_merged

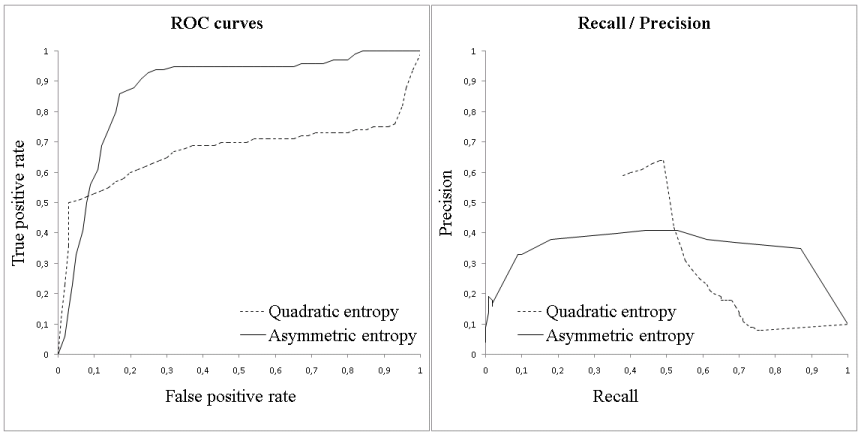

Fig. 5. Results for Satimage

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.

An important issue with asymmetric criterion is how can we determine the "most" uncertain reference distribution W? When the probability of each class

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

is known, it is consistent to use these probabilities. Otherwise, we could estimate them from the overall class frequencies in the learning dataset. For our empirical experimentation, we set this distribution once and for all. 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 [17], 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 2, 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, [18]) to find the best rate, or the Pragma criterion [10]. The choice of a rule will allow us to precisely quantify the use of an asymmetric criterion.

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 departures 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 works.

References

1. Provost, F.: Learning with imbalanced data sets. Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets (2000)
2. Barandela, R., Sánchez, J.S., García, 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. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
5. Zighed, D.A., Marcellin, S., Ritschard, G.: Mesure d'entropie asymétrique et consistante. In: EGC. (2007) 81–86
6. Rényi, A.: On measures of entropy and information. 4th Berkely Symp. Math. Statist. Probability **1** (1960) 547–561
7. Aczel, J., Daroczy, Z.: On measures of information and their characterizations. (1975)
8. Zighed, D., Rakotomalala, R.: Graphe d'induction Apprentissage et Data Mining. Hermès, Paris (2000)
9. Lallich, S., Lenca, P., Vaillant, B.: Construction d'une entropie décentrée pour l'apprentissage supervisé. 3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique (2007) 45–54
10. Thomas, J., Jouve, P.E., Nicoloyannis, N.: Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés. 3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique (2007)
11. 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
12. 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
13. Egan, J.: Signal detection theory and roc analysis. Series in Cognition and Perception (1975)
14. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letter **27**(8) (2006) 861–874
15. Hettich, S., Bay, S.D.: The uci kdd archive (1999)
16. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. (2004)
17. Chai, X., Deng, L., Yang, Q., Ling: Test-cost sensitive naive bayes classification. ICDM (2005)
18. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34(1)** (2002) 1–47