A Comparison of Different Off-Centered Entropies to Deal with Class Imbalance for Decision Trees

Philippe Lenca[1], Stéphane Lallich[2],
Thanh-Nghi Do[3], and Nguyen-Khang Pham[4]

[1] Institut TELECOM, TELECOM Bretagne, Lab-STICC, Brest, France
philippe.lenca@telecom-bretagne.eu
[2] Université Lyon, Laboratoire ERIC, Lyon 2, Lyon, France
stephane.lallich@univ-lyon2.fr
[3] INRIA Futurs/LRI, Université de Paris-Sud, Orsay, France
dtnghi@lri.fr
[4] IRISA, Rennes, France
pnguyenk@irisa.fr

Abstract. In data mining, large differences in prior class probabilities known as the class imbalance problem have been reported to hinder the performance of classifiers such as decision trees. Dealing with imbalanced and cost-sensitive data has been recognized as one of the 10 most challenging problems in data mining research. In decision trees learning, many measures are based on the concept of Shannon's entropy. A major characteristic of the entropies is that they take their maximal value when the distribution of the modalities of the class variable is uniform. To deal with the class imbalance problem, we proposed an off-centered entropy which takes its maximum value for a distribution fixed by the user. This distribution can be the a priori distribution of the class variable modalities or a distribution taking into account the costs of misclassification. Others authors have proposed an asymmetric entropy. In this paper we present the concepts of the three entropies and compare their effectiveness on 20 imbalanced data sets. All our experiments are founded on the C4.5 decision trees algorithm, in which only the function of entropy is modified. The results are promising and show the interest of off-centered entropies to deal with the problem of class imbalance.

Keywords: Decision trees, Shannon entropy, Off-centered entropies, Imbalance class.

1 Class Imbalance Problem

In supervised learning, the data set is said to be imbalanced if the class prior probabilities are highly unequal. In the case of two-class problems, the larger class is called the majority class and the smaller the minority class. Real-life two-class problems have often minority class prior under 0.10 (e.g. fraud detection, medical diagnostic or credit scoring). In such a case the performances of data mining algorithms are lowered, especially the error rate corresponding to the

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 634–643, 2008.

minority class, even though the minority class corresponds to positive cases and the cost of misclassifying the positive examples is higher than the cost of misclassifying the negative examples. This problem gave rise to many papers, from which one can cite papers from [1], [2] and [3]. Dealing with imbalanced and cost-sensitive data has been recognized as one of the 10 most challenging problems in data mining [4]. As summarized by the review papers of [5], [6] and [7] or by the very comprehensive papers of [8] and [9], solutions to the class imbalance problems were proposed both at the data and algorithmic level.

At the data level, these solutions change the class distribution. They include different forms of re-sampling, such that over-sampling [3] [10] or under-sampling [11], on a random or a directed way. A comparative study using C4.5 [12] decision tree show that under-sampling beat over-sampling [13]. At the algorithmic level, a first solution is to re-balance the error rate by weighting each type of error with the corresponding cost [14]. A study of the consistency of re-balancing costs, for misclassification costs and class imbalance, is presented in [15]. For a comparison of a cost sensitive approach and a sampling approach one can see for example [16]. In decision trees learning, other algorithmic solutions consist in adjusting the probabilistic estimates at the tree leaf or adjusting the decision thresholds. [17] propose to use a criterion of minimal cost, while [18] explore efficient pre-pruning strategies for the cost-sensitive decision tree algorithm to avoid overfitting. At both levels, [19] studied three issues (quality of probabilistic estimates, pruning, and effect of preprocessing the imbalanced data set), usually considered separately, concerning C4.5 decision trees and imbalanced data sets.

Our contribution belongs to the second category. We propose to replace the entropy used in tree induction algorithms by an off-centered entropy. That is to say that we work at the split level of decision trees learning taking into account an entropy criterion. The rest of the paper is organized as follows. In Section 2, we first review splitting criteria based on Shannon's entropy. We first recall basic considerations on Shannon's entropy and then briefly present our off-centered entropy and the asymmetric entropy. Then, we compare the entropies' performance on 20 imbalanced data sets in Section 3. Finally, Section 4 draws conclusions and suggests future work.

2 From Shannon's Entropy to Non-centered Entropies

In this section we first recall basic considerations on Shannon's entropy and then present the two families of non-centered entropies. For both of them we mainly present the boolean case and mention the results in the general case. Experiments presented in Section 3 are done in the boolean case.

2.1 Usual Measures Based on Shannon's Entropy

In supervised learning of induction tree on categorical variables, many learning algorithms use predictive association measures based on the entropy proposed by Shannon [20]. Let us consider a class variable Y having q modalities,

$p = (p_1, \ldots, p_q)$ be the vector of frequencies of Y, and a categorial predictor X having k modalities. The joint relative frequency of the couple (x_i, y_j) is denoted $p_{ij}, i = 1, \ldots k; j = 1, \ldots q$. What is more, we denote by $h(Y) = -\sum_{j=1}^{q} p_{.j} \log_2 p_{.j}$ the a priori Shannon's entropy of Y and by $h(Y/X) = E(h(Y/X = x_i))$ the conditional expectation of the entropy of Y with respect to X.

Shannon's entropy, is a real positive function of $p = (p_1, \ldots, p_q)$ to [0..1], verifying notably interesting properties for machine learning purposes:

1. **Invariance by permutation of modalities:** $h(p)$ does not change when the modalities of Y are permuted;
2. **Maximality:** the value of $h(p)$ reaches its maximum $\log_2(q)$ when the distribution of Y is uniform, i.e. each modality of Y has a frequency of $1/q$;
3. **Minimality:** the value of $h(p)$ reaches its minimum 0 when the distribution of Y is sure, centered on one modality of Y, the others being of null frequency;
4. **Strict concavity:** the entropy $h(p)$ is a strictly concave function.

Amongst the measures based on Shannon's entropy, particularly studied in by [21] and [22], we especially wish to point out:

- the entropic gain [23], which values $h(Y) - h(Y/X)$;
- the u coefficient [24] is the relative gain of Shannon's entropy i.e. the entropic gain normalized on the a priori entropy of Y, and values $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- the gain-ratio [12] which relates the entropic gain of X to the entropy of X, rather than to the a priori entropy of Y in order to discard the predictors having many modalities. It values $\frac{h(Y)-h(Y/X)}{h(X)}$;
- the Kvalseth coefficient [25], which normalizes the entropic gain by the mean of the entropies of X and Y. It then values $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$.

The peculiarity of these coefficients is that Shannon's entropy of a distribution reaches its maximum when this distribution is uniform. Even though it is the entropic gain with respect to the a priori entropy of Y which is used in the numerator part of the previously mentioned coefficients, the entropies of Y and $Y/X = x_i$ used in this gain are evaluated on a scale for which the reference value (maximal entropy) corresponds to the uniform distribution of classes. The behavior of Shannon's entropy is illustrated in Fig. 1 in the boolean case.

It would seem more logical to evaluate directly the entropic gain through the use of a scale for which the reference value would correspond to the a priori distribution of classes. The above-mentioned characteristic of the coefficients based on the entropy is particularly questionable when the classes to be learned are highly imbalanced in the data, or when the classification costs differ largely.

2.2 Off-Centered Entropy

The construction of an off-centered entropy principle is sketched out in the case of a boolean class variable in [26] and [27]. In these previous works we proposed a parameterized version of several statistical measures assessing the interest of association rules and constructed an off-centered entropy.

Let us consider a class variable Y made of $q = 2$ modalities. The frequencies distribution of Y for the values 0 and 1 is noted $(1-p, p)$. We wish to define an off-centered entropy associated with $(1-p, p)$, noted $\eta_\theta(p)$, which is maximal when $p = \theta$, θ being fixed by the user and not necessarily equal to 0.5 (in the case of a uniform distribution). In order to define the off-centered entropy, following the proposition described in [26], we propose that the $(1-p, p)$ distribution should be transformed into a $(1-\pi, \pi)$ distribution such that: π increases from 0 to 1/2 when p increases from 0 to θ, and π increases from 1/2 to 1 when p increases from θ to 1. By looking for an expression of π as $\pi = \frac{p-b}{a}$, on both intervals $0 \leq p \leq \theta$ and $\theta \leq p \leq 1$, we obtain: $\pi = \frac{p}{2\theta}$ if $0 \leq p \leq \theta$, $\pi = \frac{p+1-2\theta}{2(1-\theta)}$ if $\theta \leq p \leq 1$.

To be precise, the thus transformed frequencies should be denoted as $1 - \pi_\theta$ and π_θ. We will simply use $1 - \pi$ and π for clarity reasons. They do correspond to frequencies, since $0 \leq \pi \leq 1$. The off-centered entropy $\eta_\theta(p)$ is then defined as the entropy of $(1-\pi, \pi)$: $\eta_\theta(p) = -\pi \log_2 \pi - (1-\pi) \log_2(1-\pi)$.

With respect to the distribution $(1-p, p)$, clearly $\eta_\theta(p)$ is not an entropy strictly speaking. Its properties must be studied considering the fact that $\eta_\theta(p)$ is the entropy of the transformed distribution $(1-\pi, \pi)$, *i.e.* $\eta_\theta(p) = h(\pi)$. The behavior of this entropy is illustrated in Fig. 1 for $\theta = 0.2$.

The off-centered entropy preserves various properties of the entropy, among those studied in particular by [28] in a data mining context. Those properties are easy to prove since $\eta_\theta(p)$ is defined as an entropy on π and thus possess such characteristics. It can be noticed that $\eta_\theta(p)$ is maximal for $p = \theta$ i.e. for $\pi = 0.5$. Invariance by permutation of modalities property is of course voluntarily abandoned. Proofs are given in detail in [29].

Following a similar way as in the boolean case we then extended the definition of the off-centered entropy to the case of a variable Y having q modalities, $q > 2$ [29,30]. The off-centered entropy for a variable with $q > 2$ modalities is the defined by $\eta_\theta(p) = h(\pi^*)$ where: $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^{q} \pi_j}$ (in order to satisfy the normalization property), $0 \leq \pi_j \leq 1$, $\sum_{j=1}^{q} \pi_j = 1$ (π_j should be analogous to frequencies), $\pi_j = \frac{p_j}{q\theta_j}$ if $0 \leq p_j \leq \theta_j$, $\pi_j = \frac{q(p_j-\theta_j)+1-p_j}{q(1-\theta_j)}$ if $\theta_j \leq p_j \leq 1$.

2.3 Off-Centered Generalized Entropies

Shannon's entropy is not the only diversity or uncertainty function usable to build coefficients of predictive association. [31] already proposed a unified view of the three usual coefficients (the λ of Guttman, the u of Theil and the τ of Goodman and Kruskal), under the name of *Proportional Reduction in Error* coefficient. In a more general way we built the *Proportional Reduction in Diversity* coefficients, which are the analogue of the standardized gain when Shannon's entropy is replaced by whichever concave function of uncertainty [32].

One of the particularities of the off-centering we here propose, compared to the approach proposed by [33] is that rather than defining a single off-centered entropy, it adapts to whichever kind of entropy. We thus propose a decentring framework that one can apply to any measure of predictive association based on a gain of uncertainty [30].

2.4 Asymmetric Entropy

With an alternative goal, directly related to the construction of a predictive association measure, especially in the context of decision trees, [34] proposed a consistent and asymmetric entropy for a boolean class variable. This measure is asymmetric in the sense that one may choose the distribution for which it will reach its maximum; and consistent since it takes into account n, the size of the sampling scheme. They preserve the *strict concavity* property but alter the *maximality* one in order to let the entropy reach its maximal value for a distribution chosen by the user (*i.e.* maximal for $p = \theta$, where θ is fixed by the user). This implies revoking the *invariance by permutation of modalities.* They propose: $h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p+\theta^2}$. It can be noticed that for $\theta = 0.5$, this asymmetric entropy corresponds to the quadratic entropy of Gini. The behavior of this entropy is illustrated in Fig. 1 for $\theta = 0.2$.

In [33], the same authors extend their approach to the situation where the class variable has $q > 2$ modalities. What is more, since one may only make an estimation of the real distribution $(p_j)_{j=1,\ldots,q}$ with an empirical distribution $(f_j)_{j=1,\ldots,q}$, they wish that for same values of the empirical distribution, the value of the entropy should decrease as n rises (property 5, a new property called *consistency*). They thus are led to modify the third property (*minimality*) in a new property 3′ (*asymptotic minimality*): the entropy of a sure variable is only required to tend towards 0 as $n \to \infty$. In order to comply with these new properties, they suggest to estimate the theoretical frequencies p_j by their Laplace estimator, $\widehat{p_j} = \frac{nf_j+1}{n+q}$. They thus propose a consistent asymmetric entropy as: $h_\theta(p) = \sum_{j=1}^{q} \frac{\widehat{p_j}(1-\widehat{p_j})}{(1-2\theta_j)\widehat{p_j}+\theta_j^2}$.

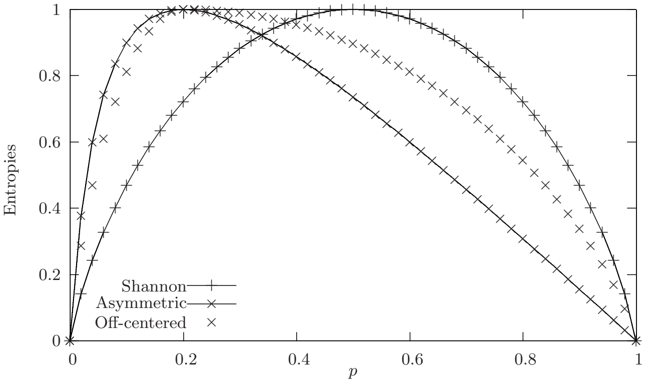

Fig. 1. Off-centered, asymmetric and Shannon's entropies

3 Experiments with More or Less Imbalanced Data Sets

In our experiments, we compare the behaviors of decision tree algorithms to classify imbalanced data sets using our proposed off-centered entropy OCE, the Shannons entropy SE and the asymmetric entropy AE. To achieve the evaluation we added OCE and AE to the decision tree algorithm C4.5 [12]. In these experiments, in each node the distribution for which OCE and AE are maximal is the a priori distribution of the class variable in the considered node.

The experimental setup used the 20 data sets described in Table 1 (column 1 indicates the data set name, the numbers of cases and of attributes), where the first twelve ones are from the UCI repository [35], the next six are from the Statlog repository [36], the following data set is from the DELVE repository (http://www.cs.toronto.edu/∼delve/), while the last one is from [37].

In order to evaluate the performance of the considered entropies for classifying imbalanced data sets, we pre-processed multi-class (more than two classes, denoted by an asterisk) data sets as two-class problems. The columns 2 and 3 of Table 1 show how we convert multi-class to minority and majority classes. For example, with the OpticDigits data set, the digit 0 is mapped to the minority class (10%) and the remaining data are considered as the majority class (90%). For the 20-newsgroup collection, we pre-processed the data set by representing each document as a vector of words. With a feature selection method which uses mutual information, we get a binary data set of 500 dimensions (words).

The test protocols are presented in the column 4 of Table 1. Some data sets are already divided in training set (trn) and testing set (tst). If the training set and testing set are not available then we used cross-validation protocols to evaluate the performance, else k-fold cross validation is used. With a data set having less than 300 data points, the test protocol is leave-one-out cross-validation (loo). It involves using a single data point of the data set as the testing data and the remaining data points as the training data. This is repeated such that each data point in the data set is used once as testing data. With a data set having more than 300 data points, k-fold cross-validation is used to evaluate the performance. In k-fold cross-validation, the data set is partitioned into k folds. A single fold is retained as the validation set for testing, and the remaining k-1 folds are used as training data. The cross-validation process is then repeated k times. The k results are then averaged. The columns 5 to 9 of Table 1 present the results according to each entropy in terms of tree size, global error rate, error rate on the minority class and on the majority class (best results are in bold). The synthetic comparisons two by two are presented in Table 2.

For these first comparisons, we recall that the rule of prediction is the majority rule. The definition of another rule of prediction, better adapted to non-centered entropies, is one of the enhancements which we intend to accomplish.

We can conclude that the non-centered entropies, particularly the off-centered entropy, outperform the Shannon's entropy. These both entropies significantly improve the MinClass accuracy, without penalizing the MajClass accuracy, where MinClass (MajClass) accuracy is the proportion of true results in the minority (majority) class.

Table 1. Experiments on 20 imbalanced data sets

Base	Class Min.	Class Maj.	Valid.	Method	Tree size	Acc.	MinClass acc.	MajClass acc.
Opticdigits*	10%(0)	90%(rest)	trn-tst	SE	27	99.39	96.63	99.69
5620				AE	**21**	**99.83**	**100.00**	**99.81**
64				OCE	**21**	99.67	99.44	99.69
Tictactoe	35%(1)	65%(2)	10-fold	SE	**69**	93.33	87.50	**96.49**
958				AE	89	93.65	89.52	95.82
9				OCE	89	**94.17**	**90.43**	96.15
Wine*	27%(3)	73%(rest)	loo	SE	**5**	95.51	89.58	**97.69**
178				AE	**5**	**97.19**	**95.83**	**97.69**
13				OCE	**5**	**97.19**	**95.83**	**97.69**
Adult	24%(1)	76%(2)	trn-tst	SE	123	**86.25**	60.85	**94.11**
48842				AE	171	85.67	60.02	93.61
14				OCE	**107**	85.70	**61.61**	93.15
20-newsgrp*	5%(1)	95%(rest)	3-fold	SE	**9**	98.59	73.31	**99.95**
20000				AE	13	**98.65**	**74.49**	**99.95**
500				OCE	13	**98.65**	**74.49**	**99.95**
Breast Cancer	35%(M)	65%(B)	10-fold	SE	18	94.04	90.43	96.31
569				AE	**11**	**94.39**	90.40	**96.90**
30				OCE	13	93.33	**90.45**	95.20
Letters*	4%(A)	96%(rest)	3-fold	SE	**67**	**99.47**	91.48	**99.81**
20000				AE	99	99.35	90.00	99.75
16				OCE	105	99.44	**92.59**	99.73
Yeast*	31%(CYT)	69%(rest)	10-fold	SE	52	71.76	47.95	82.66
1484				AE	65	71.82	**48.82**	82.26
8				OCE	**34**	**72.34**	47.00	**84.02**
Connect-4*	10%(draw)	90%(rest)	3-fold	SE	4141	83.25	57.02	91.72
67557				AE	**4013**	83.46	57.59	91.81
42				OCE	4037	**84.07**	**60.09**	**91.82**
Glass*	33%(1)	67%(rest)	loo	SE	39	77.10	72.41	80.32
214				AE	23	78.97	72.86	81.94
9				OCE	**21**	**86.45**	**78.57**	**90.28**
Spambase	40%(spam)	60%(rest)	10-fold	SE	250	93.00	90.94	94.31
4601				AE	269	93.22	**91.52**	94.28
57				OCE	**225**	**93.35**	91.21	**94.67**
Ecoli*	15%(pp)	85%(rest)	10-fold	SE	**11**	94.55	74.68	**98.19**
336				AE	14	94.24	76.50	97.43
7				OCE	**11**	**95.45**	**81.93**	97.80
Pima	35%(1)	65%(2)	10-fold	SE	25	74.94	62.79	81.42
768				AE	**20**	**75.71**	**64.30**	**81.82**
8				OCE	**20**	75.19	63.15	81.62
German	30%(1)	70%(2)	10-fold	SE	**39**	74.27	40.00	**88.36**
1000				AE	40	73.54	40.07	86.95
20				OCE	43	**74.48**	**44.40**	86.45
Shuttle*	20%(rest)	80%(1)	trn-tst	SE	27	**99.99**	99.93	**100.00**
58000				AE	19	99.80	99.90	**100.00**
9				OCE	**11**	**99.99**	**99.97**	**100.00**
Segment*	14%(1)	86%(rest)	10-fold	SE	**7**	99.22	95.78	99.79
2310				AE	18	**99.31**	95.91	**99.85**
19				OCE	19	**99.31**	**96.75**	99.75
Satimage*	24%(1)	90%(rest)	trn-tst	SE	99	97.35	94.36	98.25
6435				AE	103	**98.00**	**96.10**	98.57
36				OCE	**93**	97.95	95.23	**98.77**
Vehicle*	24%(van)	76%(rest)	10-fold	SE	41	94.81	88.49	95.70
846				AE	**31**	94.94	**90.66**	96.33
18				OCE	32	**95.18**	88.95	**97.10**
Splice*	25%(EI)	75%(rest)	10-fold	SE	72	96.37	92.74	**97.62**
3190				AE	62	**96.40**	93.23	97.50
60				OCE	**24**	**96.40**	**93.69**	97.33
All-Aml	35% (AML)	65%(ALL)	loo	SE	**3**	**91.18**	**92.86**	**90.00**
72				AE	**3**	**91.18**	**92.86**	**90.00**
7129				OCE	**3**	**91.18**	**92.86**	**90.00**

Table 2. Comparison of Shannon entropy (SE), Off-centered entropy (OCE) and Asymmetric entropy (AE)

OCE vs. SE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (OCE-SE)	-9.900	0.76%	1.94%	0.44%
Mean Std. dev. (OCE-SE)	6.318	0.47%	0.53%	0.53%
Student ratio	-1.567	1.621	3.673	0.830
p-value (Student)	Non sign.	Non sign.	0.0016	Non sign.
OCE wins	12	16	18	7
Exaequo	3	1	1	5
SE wins	5	3	1	8
p-value (sign test)	Non sign.	0.0044	0.0000	Non sign.

AE vs. SE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (AE-SE)	-1.750	0.25%	1.04%	-0.01%
Mean Std. dev. (AE-SE)	7.500	0.14%	0.37%	0.14%
Student ratio	-0.233	1.746	2.808	-0.048
p-value (Student)	Non sign.	0.0970	0.0112	Non sign.
AE wins	8	14	15	8
Exaequo	2	1	1	4
SE wins	10	5	4	8
p-value (sign test)	Non sign.	Non sign.	0.0192	Non sign.

OCE vs. AE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (OCE- AE)	-8.150	0.51%	0.90%	0.45%
Mean Std. dev. (OCE- AE)	4.563	0.38%	0.49%	0.44%
Student ratio	-1.786	1.330	1.846	1.014
p-value (Student)	0.0901	0.1991	0.0805	0.3234
OCE wins	8	11	11	8
Exaequo	6	5	3	4
AE wins	6	4	6	8
p-value (sign test)	Non sign.	Non sign.	Non sign.	Non sign.

Indeed, compared to Shannon's entropy SE, the off-centered entropy OCE improves the MinClass accuracy 18 times out of 20, with 1 defeat and 1 equality, which corresponds to a p-value of 0.0000. The corresponding average gain in accuracy is close to 0.02 (p-value = 0.0016 according to a paired t-test). The accuracy of the MajClass is not significantly modified, but the global accuracy is improved 16 times out of 20, with 3 defeats and 1 equality (p-value = 0.0044), while the average corresponding gain is close to 0.008. Moreover, the trees provided by OCE have often a more reduced size, but this reduction is not significant.

The asymmetric entropy AE gives slightly less significant results when compared to Shannon's entropy SE. It improves 15 times out of 20 the MinClass accuracy (p-value = 0.0192), with an average gain close to 0.01 (p-value = 0.0112). However, the improvement of the global accuracy is not significant: AE wins 14 times out of 20, with 1 equality and 5 defeats, while the increase of the global accuracy is only 0.002. In the same way, the performance for the MajClass accuracy is comparable (AE wins 8 times, SE wins 8 times, and 4 equalities). Furthermore, for the size of the tree, the performance is also comparable (AE wins 8 times, SE wins 10 times, and 2 equalities).

When comparing the two non-centered entropies OCE and AE, one can observe a slight but not significant superiority of the off-centered entropy OCE for each criterion. Particularly, a gain of 1 point on the MinClass error rate and 0.5 point on the total error rate must be noticed.

4 Conclusion and Future Works

In order to deal with imbalanced classes, we proposed an off-centered split function for learning induction trees. It has the characteristic to be maximum for the distribution a priori of the class in the node considered. We then compare, in the boolean case on 20 imbalanced data bases, the performances of our entropy with the entropy of Shannon and an asymmetric entropy. All our experiments are founded on C4.5 decision trees algorithm, in which only the entropy is modified. Compared to Shannon's entropy both non-centered entropies, significantly improve the minority class accuracy, without penalizing the majority one. Our off-centered entropy is slightly better than the asymmetric one, but this is not statistically significant. However one major advantage of our proposal is that it can be applied to any kind of entropy, for example to the quadratic entropy of Gini used in the CART algorithm [38]. We plan to improve the pruning scheme and the criterion to affect a class to a leaf. Indeed, these two criteria such as defined in C4.5, do not well support the recognition of the minority class. We then can hope for an improvement of our already good results. It could be also valuable to take into account a cost-sensitive matrix.

References

1. Japkowicz, N. (ed.): Learning from Imbalanced Data Sets/AAAI (2000)
2. Chawla, N., Japkowicz, N., Kolcz, A. (eds.): Learning from Imbalanced Data Sets/ICML (2003)
3. Chawla, N., Japkowicz, N., Kolcz, A. (eds.): Special Issue on Class Imbalances. SIGKDD Explorations, vol. 6 (2004)
4. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making 5(4), 597–604 (2006)
5. Japkowicz, N.: The class imbalance problem: Significance and strategies. In: IC-AI, pp. 111–117 (2000)
6. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429–450 (2002)
7. Visa, S., Ralescu, A.: Issues in mining imbalanced data sets - A review paper. In: Midwest AICS Conf., pp. 67–73 (2005)
8. Weiss, G.M., Provost, F.: The effect of class distribution on classifier learning. TR ML-TR 43, Department of Computer Science, Rutgers University (2001)
9. Weiss, G.M., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. J. of Art. Int. Research 19, 315–354 (2003)
10. Liu, A., Ghosh, J., Martin, C.: Generative oversampling for mining imbalanced datasets. In: DMIN, pp. 66–72 (2007)
11. Kubat, M., Matwin, S.: Addressing the curse of imbalanced data sets: One-sided sampling. In: ICML, pp. 179–186 (1997)
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
13. Drummond, C., Holte, R.: C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In: Learning from Imbalanced Data Sets/ICML (2003)
14. Domingos, P.: Metacost: A general method for making classifiers cost sensitive. In: KDD, pp. 155–164 (1999)

15. Zhou, Z.H., Liu, X.Y.: On multi-class cost-sensitive learning. In: AAAI, pp. 567–572 (2006)
16. Weiss, G.M., McCarthy, K., Zabar, B.: Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In: DMIN, pp. 35–41 (2007)
17. Ling, C.X., Yang, Q., Wang, J., Zhang, S.: Decision trees with minimal costs. In: ICML (2004)
18. Du, J., Cai, Z., Ling, C.X.: Cost-sensitive decision trees with pre-pruning. In: Kobti, Z., Wu, D. (eds.) Canadian AI 2007. LNCS (LNAI), vol. 4509, pp. 171–179. Springer, Heidelberg (2007)
19. Chawla, N.: C4.5 and imbalanced datasets: Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In: Learning from Imbalanced Data Sets/ICML (2003)
20. Shannon, C.E.: A mathematical theory of communication. Bell System Technological Journal (27), 379–423, 623–656 (1948)
21. Wehenkel, L.: On uncertainty measures used for decision tree induction. In: IPMU, pp. 413–418 (1996)
22. Loh, W.Y., Shih, Y.S.: Split selection methods for classification trees. Statistica Sinica 7, 815–840 (1997)
23. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
24. Theil, H.: On the estimation of relationships involving qualitative variables. American Journal of Sociology (76), 103–154 (1970)
25. Kvalseth, T.O.: Entropy and correlation: some comments. IEEE Trans. on Systems, Man and Cybernetics 17(3), 517–519 (1987)
26. Lallich, S., Vaillant, B., Lenca, P.: Parametrised measures for the evaluation of association rule interestingness. In: ASMDA, pp. 220–229 (2005)
27. Lallich, S., Vaillant, B., Lenca, P.: A probabilistic framework towards the parameterization of association rule interestingness measures. Methodology and Computing in Applied Probability 9, 447–463 (2007)
28. Zighed, D.A., Rakotomalala, R.: Graphes d'Induction – Apprentissage et Data Mining. Hermes (2000)
29. Lallich, S., Vaillant, B., Lenca, P.: Construction d'une entropie décentrée pour l'apprentissage supervisé. In: QDC/EGC 2007, pp. 45–54 (2007)
30. Lallich, S., Lenca, P., Vaillant, B.: Construction of an off-centered entropy for supervised learning. In: ASMDA, p. 8 (2007)
31. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, i. JASA I(49), 732–764 (1954)
32. Lallich, S.: Mesure et validation en extraction des connaissances à partir des données. In: Habilitation à Diriger des Recherches, Université Lyon 2, France (2002)
33. Zighed, D.A., Marcellin, S., Ritschard, G.: Mesure d'entropie asymétrique et consistante. In: EGC, pp. 81–86 (2007)
34. Marcellin, S., Zighed, D.A., Ritschard, G.: An asymmetric entropy measure for decision trees. In: IPMU, pp. 1292–1299 (2006)
35. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998)
36. Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds.): Machine Learning, Neural and Statistical Classification. Ellis Horwood (1994)
37. Jinyan, L., Huiqing, L.: Kent ridge bio-medical data set repository. Technical report (2002)
38. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International (1984)