Studying the Behavior of Generalized Entropy in Induction Trees Using a M-of-N Concept

R. Rakotomalala, S. Lallich and S. Di Palma

ERIC Laboratory - University of Lyon 2
e-mail:{rakotoma,lallich,sdipalma}@univ-lyon2.fr

Abstract

This paper study splitting criterion in decision trees using three original points of view. First we propose a unified formalization for association measures based on entropy of type beta. This formalization includes popular measures such as Gini index or Shannon entropy. Second, we generate artificial data from M-of-N concepts whose complexity and class distribution are controlled. Third, our experiment allows us to study the behavior of measures on datasets of growing complexity. The results show that the differences of performances between measures, which are significant when there is no noise in the data, disappear when the level of noise increases.

1 Introduction

Induction tree methods such as CART [1], C4.5 [10] or generalized approach like SIPINA [14] know a great success in data mining research because they are very fast and easy to use. The learning goal is to produce subgroups as homogeneous as possible considering a particular attribute, called the class. The induction algorithm they use to achieve this goal is simple : split each node of the tree, which represents a subset of the whole population, using a predictive attribute of the learning set until a stopping rule is activated. The selection of the predictive attribute relies only on a splitting measure that allows to order attributes according to their contribution to predicting the value of the class attribute. Many works have been devoted to this crucial element of induction graph methods [11]: some try to classify the measures used in practice [12] while others compare their performances on benchmark databases [2]. The behavior of these measures in a learning process remains however largely unknown, notably because the studies often use databases whose characteristics are not specified so that results are finally only validated on studied databases.

In this paper, we deepen the study of splitting measures from three original point of view. First we adopt an unified formalization of splitting measures which includes most existing measures such as Gini index [1] or Shannon entropy [10] by modulating a parameter β. Second, we generate artificial databases using a M-of-N concept which allows us to totally control class distribution and the complexity of the concept. Various levels and kind of noise can thus be used to study the behavior of generalized entropy. Third, we compare performances of decision trees on M-of-N concepts of growing complexity.

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 510–517, 1999.

In the following, we present M-of-N concepts and their transformation into trees. We then introduce a generalization of normalized gain entropy measure and study its behavior in induction trees algorithm using artificial datasets generated with a M-of-N concept.

2 From M-of-N Concepts to Decision Trees

The use of M-of-N concepts allows us to control the difficulty of the learning process. These concepts are especially hard to learn with decision tree [8] [9]. We note that it does not concern simply an artificial concept, M-of-N can occur in real problem [4].

We first recall the M-of-N concept definition. Let us consider N independent boolean variables with the same probability p. We call M-of-N concept, the boolean variable which takes the value 1 if and only at least M of the N variables take the value 1. For example a 2-of-3 concept, with three boolean attributes (A, B and C) is logically equivalent to $AB + AC + BC$.

2.1 Checking the Complexity of M-of-N Concepts

A M-of-N concept is a disjunction of $C_N^M = \frac{N!}{M!(N-M)!}$ conjunctions of length M. The minimal tree necessary to learn M-of-N concept is of depth N, but the number of leaves of this tree relies on M. To evaluate the complexity of M-of-N concepts, we propose to calculate the number of leaves of the logically equivalent tree. This number of leaves, denoted by F_N^M, is calculated as follow[1] :

1. We calculate the number of leaves at the level i for the minimal tree, i = 1, 2, ..., N-1. A node of the minimal tree is a leaf since the one or the other of the two exclusive conditions blow is satisfied :
   - C_1 : the node corresponds to a "1" and there are M-1 nodes above it on its branch corresponding to a "1" ;
   - C_2 : the node corresponds to a "0" and there are $(N-M)$ nodes above it on its branch corresponding to a "0" ; At the level i of the tree, there are C_{i-1}^{M-1}nodes satisfying C_1, $i \geq m$,while there are C_{i-1}^{N-M} nodes satisfying C$_2$, $i \geq n - m + 1$.
2. The total number of leaves is obtained by summing for the values of i :

$$F_N^M = \sum_{i=M}^{N} C_{i-1}^{M-1} - \sum_{i=N-M+1}^{N} C_{i-1}^{N-M} = \sum_{j=0}^{N-M} C_{M-1+j}^{M-1} - \sum_{j=0}^{M-1} C_{i-1}^{N-M}$$

$$F_N^M = C_N^{N-M} + C_N^{M-1} = C_{N+1}^M$$

where $F_N^1 = F_N^N = N + 1$. We choose to work with 3-of-N concepts, $N = 3, 4, ..., 7$, because they show a sufficient range of complexity.

[1] Furthermore, we can calculate the number of nodes necessary to M-of-N concept learning (exception of the root node) : $N_N^M = 2(F_N^M - 1)$

Table1. Values of p, F_N^M and p_{min} for "3-of-N" concepts, N = 3,...,7

N	p	F_N^3	p$_{min}$
3	0.6300	4	0.1469
4	0.4563	10	0.0615
5	0.3594	20	0.0340
6	0.2969	35	0.0215
7	0.2531	56	0.0149

2.2 Checking the Probability Distribution of the output class

We decided to learn concepts with the same class probabilities distribution in order to avoid the comparison to be altered by the nature more and less unbalanced of these distributions. We opted for the probabilities distribution (0.75 ; 0.25) whose imbalance is intermediate.

To generate our datasets, we use N independent boolean attributes which take the value 1 (TRUE) with the probability p. To calculate the resulting class distribution of the M-of-N concept, let us define K_N to be the number of variables which take the value 1 among the N attributes. Then K_N has a binomial distribution with parameters N and p. Thus, the positive class probability is

$$P(C_{M/N}=1)=P(K_N\geqslant M)=1-p'_{M-1}$$

where p'_M is the cumulated probability of the binomial distribution $B(N;p)$ for the value M.

Table 1 gives p, the number of leaves and the value of p$_{min}$(probability of the least probable rule) used in learning set size definition for $N=3,4,...,7$.

3 Association Measure based on Entropy of Type β

Let us consider a variable to be learned C, made up of K categories c$_k$, k=1, 2,..., K, and a predictive attribute X made up of L categories x$_l$, l=1, 2, ..., L. We denote by π_{kl} the joint probability of c$_k$ and x$_l$, π_{k+} and π_{+l} the marginal probabilities of C and X.

To measure the predictive association between C and X, one usually calculates the relative mean reduction of uncertainty on the distribution of C due to the knowing of X, following the P.R.E coefficients of association (proportional reduction in error) proposed by Goodman and Kruskal [3]. We propose to establish a generalized measure of association $L_\beta(C/X)$, based on the mean relative reduction of generalized entropy of the distribution of C gained when knowing X. Actually, the entropy of C can only decrease when reasoning conditionally to $X=x$, because of the concavity of the entropy of type β, $\beta>0$. We have established the generic formula of this measure, and proved its good properties. The entropy of type ß of C is defined by: $H_\beta(C)=\frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1-\sum_{k=1}^{K}\pi_{k+}^{\beta}\right)$. We obtain for the formula of L_β(C/X):

$$L_\beta(C/X) = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K} \pi_{kl}^{\beta}\pi_{+l}^{1-\beta} - \sum_{k=1}^{K}\pi_{k+}^{\beta}}{1 - \sum_{k=1}^{K}\pi_{k+}^{\beta}}$$

Because sample counts have a multinomial distribution, the corresponding proportions are asymptotically normal as well as $L_\beta(C/X)$. We have calculated its asymptotic variance by applying the delta method [6].

The specific values of ß enable us to find not only the usual measures of association based on Shannon entropy (for ß=1, by passage to the limit) and Gini quadratic entropy (for ß=2), but also the ones based on the number of categories (ß=0) or on Bhattacharya affinity (ß=0.5). Thus, by modulating the parameter ß during the experimentation, we can synthetically compare the efficiency of usual association coefficients.

4 Experiment : Concept, Learning Set Size and Noise

Taking into account our precedent paragraphs, the use of artificial datasets gives several advantages : we can control the difficulty and the probability distribution of the concept to learn, define theoretically the minimum size of the learning set and introduce controlled random noise.

4.1 Concept to Learn

In our experimentation, we study five boolean concepts of growing complexity, which are 3-of-N concepts with $N = 3, 4, ..., 7$. They are constructed by generating boolean attributes, that take the value TRUE with the probability p defined in table 2, and then applying the function to learn in order to obtain the value of the variable to predict.

These functions are disjunctive normal forms, so they are very hard to learn for an induction tree. Ten independent boolean predictive attributes were added in our datasets. Indeed, if we confine to attributes of the concept functions, they will always be selected, making believe to a fallacious robustness [2].

4.2 Size of the Learning Set

Three parameters are to be set : size n of the training set, size k of validation set and the number r of repetitions of the experiment. To test if the value of the size n of the training set interfere with the optimal value of β, we tried different sizes of the training set for each concept.

We can say that an i-rule which predict "1" for the concept has the probability $p^M(1-p)^{i-M}$, while an i-rule which predict "0" for the concept has the probability $p^{i-N+M-1}(1-p)^{N-M+1}$. Consequently, the least probable rule of the minimum tree has the probability p_{min}, given by :

$$p_{\min} = Min\left\{p^{M}(1-p)^{N-M}; p^{M-1}(1-p)^{n-M+1}\right\}$$

So, according to the results of table 1 relative to the number of rules and the probability of the least probable rule for each concept, we selected four values of n for each concept, i-e ten, fifteen, twenty and twenty five time the number of rules. Thus, the size of leaves of our trees is not lower than 5 examples $(0.1469*4*10 = 5.876)$.

Concerning the test set, we generate with the same process $k = 2000$ examples, which is sufficient to obtain an accurate estimation of the error rate [9]. At last, we choose to realize $r = 25$ repetitions of the experiment, in order to be able to test the effect of the factors on the generalization accuracy rate.

4.3 Kind and Level of Noise

To complete the study of the behavior of measures, we introduced noise in our artificial datasets. Only the class attribute was added noise according to the following random procedure : for each example, the result of the concept is noised with a probability which depends on the class distribution. Three kinds of noise (a, b, c) are studied in this paper.

We denote by π_1 and π_2 the probabilities of each value of the class, θ_1 and θ_2 their probabilities to be noised, and θ the overall probability of noise, $\theta = \theta_1\pi_1 + \theta_2\pi_2$. We recall that in our experiment, $\pi_1 = 0.75$ and $\pi_2 = 0.25$. Given that noising the data modify class distributions, π_1^* and π_2^* are the probabilities of each value of the class after noise.

(a) the occurrence of noise is the same, whatever the value taken by the class attribute. We can consider it as a reference. It occurs notably in industrial process where data are collected automatically.
$\theta_1 = \theta_2 = \theta;\ \pi_i^* = \pi_i + \theta\left(\pi_{not\ i} - \pi_i\right),\ i = 1, 2$

(b) the occurrence of noise is proportional to class distribution. For instance, in medical tests, the probability of disease is often weak and the probability of "false positive" is greater than the probability of "false negative".
$\theta_i = \frac{\theta\pi_i}{\pi_1^2+\pi_2^2};\ \pi_i^* = \pi_i + \frac{\theta\left(\pi_{not\ i}^2 - \pi_i^2\right)}{\pi_1^2+\pi_2^2},\ i = 1, 2$

(c) the occurrence of noise is inversely proportional to the class distribution. This problem is very hard to learn when we have a very unbalanced class distribution. Indeed if noise is concentrated on the rare value, it is very difficult to exceed the simple classifier concluding always to the most request value in the learning set.
$\theta_i = \frac{\theta}{2\pi_i};\ \pi_i^* = \pi_i, i = 1, 2$

We introduce noise in our datasets with the following procedure : for each kind of noise, for each overall level of noise θ (0, 0.05, 0.10, 0.20), we calculate the probability of the examples related to each value of the concept to be modified. Then the examples are modified according to this probability. Each example to be modified takes the alternative class value.

5 Results and comments

To check the behavior of generalized entropy in induction tree algorithm, we use the popular C4.5 algorithm [10] where we replace the splitting criterion ”gain ratio” with entropy of type β. All others features are the same, in particular we expand the maximum trees before pruning them using the pessimistic error rate. Thus, differences between trees rely only on the value of β for the generalized entropy.

To evaluate the performance of measures, we use the generalization error rate. Comparisons focus on values of $\beta = 1.0, 1.5, 2.0, 3.0, 5.0$. In the case of each concept we have performed an ANOVA (Analysis of variance) in order to test the main effects and the interactions of the different factors on the error rate. Because we have a very big number of tests, we consider a result to be a significant one only if its p-value is lower than 0.001 $(***)$ or comprised between 0.001 and 0.01 $(**)$.

We distinguish two kinds of results, firstly those concerning the case of no noise in generated dataset used for learning, secondly those concerning the case of noised data.

5.1 Without noise

As it was foreseeable, the size n of the training set is significant $(***)$ for each concept M-of-N, $N = 3, 4, ..., 7$. If the size of the learning set is sufficient, trees are able to learn our function. We note that we do not learn always the right concept. In fact the error rate increases when the ratio size of learning set and concept complexity decreases.

The value of β is more and more significant as the complexity of the M-of-N concept increases: no significant for $N = 3, 4, 5$, very significant $(**)$ for $N = 6$ and very highly significant $(***)$ for $N = 7$. Table 2 shows that when the complexity increases, the best values of β are at the opposite extremes ($\beta = 1$ and $\beta = 5$) and the performance of the intermediate values of β (e.g. $\beta = 1.5, 2, 3$) deteriorates. Perhaps we can find explanation of the effectiveness of the measure with their empirical variance : when the variance of the measure is low, it can be better to choose the best one among candidate attributes, especially for rejecting noisy attributes. Unfortunately, the study of the variance behavior is very hard here because it relies on β, conditional and unconditional distribution, but also on the number of values of the class and the splitting attributes (for the special case of boolean concept, we would be allowed to study a 2×2 cross-tabulation).

This first result is very interesting but we can ask to know if it is useful on data mining problem. We note that even if the differences is statistically significant, they were not practically significant. For instance, for the 3-of-7 concept, errors vary from 0.0194 ($\beta = 3$) to 0.0172 ($\beta = 1$). It is obvious that this improvement will not be useful in real problem. A further study on others concepts will be necessary to confirm or reject the weakness of these differences but as we see below, on real dataset which are naturally noisy, we wonder if it is really necessary to search the best measure (the best parameter β) for a problem.

Table2. Error rate (x 10000) of decision trees built on data without noise

N \ β	1	1.5	2	3	5
3	18	15	**9**	18	14
4	106	104	99	**90**	92
5	**136**	144	159	159	**134**
6	**165**	174	186	184	**156**
7	**172**	181	192	194	**175**

5.2 With Noise

At the opposite, in case of noisy data, all ANOVAs give similar results for p-value of main factor and their interactions, says :

- size of learning set, the level of noise and their interaction are very highly significant ($***$) whatever the complexity of the concept M-of-N. This interaction relies on the reduction of the action of learning set size when the level of noise increases.
- neither the value of β, nor its interactions with the other factors are significant. Especially, there is no interaction of the kind of noise with the level of β whatever the level of noise.

An important result of our study is to note that statistical differences between error rates according to the value of β observed in table 2 disappear when we use noisy data.

6 Conclusion

Our paper presents an unified framework for splitting measures which generalizes the standard one. This metric depends on a β parameter that we can vary to obtain famous measures such as Shannon entropy or Gini index. Normalized measures such as Gain ratio or Mantaras distances can also be deduced[7] [10]. Parametric indicators, especially the asymptotic variance, were calculated.

We evaluate the behavior of this generalized entropy measure in decision tree algorithm according to the β parameter. Our originality beside previous studies [2] [13] is to use concept M-of-N which provides a scale of functions whose we control the growing complexity and the class distribution.

The first result of our work is that the differences of performances between measures, which are statistically significant when there is no noise in datasets, disappear when the level of noise increases. This could explain the conclusions of many authors who note that measures influence the size of trees rather than their performance [5]. Indeed, most of them use real datasets which are naturally noisy, they don't control class distribution or concept complexity. Our study, using synthetic dataset, is more powerful to detect differences between measures behavior.

However, and this is our second main results, we note that this differences are statistically significant but not practically significant, even on artificial dataset built from boolean concept. We expect that for the most part of real datasets which are often more or less noisy, all measures issued from generalized entropy give a good approach to specialize induction trees.

References

1. L. Breiman, J.H. Friedman, R.A. Olshen et C.J. Stone. *Classification and Regression Trees.* California : Wadsworth International, 1984.
2. W. Buntine et T. Niblett. A further comparison of splitting rules for decision tree induction. *Machine Learning*, 8:75–85, 1992.
3. L. A. Goodman et W. H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 37:54–115, 1954.
4. L.C. Kingsland. The evaluation of medical expert system : Experience with the ai/rheum knowledge-based consultant in rheumatology. In *Proceedings of the Ninth Annual Symposium on Computer Applications in Medical Care*, 1985.
5. I. Kononenko. On biases in estimating multi-valued attributes. In *Proc. Int. Joint Conf. On Artificial Intelligence IJCAI'95*, pages 1034–1040, 1995.
6. S. Lallich. Concept de diversite et association predictive. In *Proceedings of XXXIemes Journees de Statistique*, pages 673–676, May 1999.
7. R.L. De Mantaras. A distance-based attributes selection measures for decision tree induction. *Machine Learning*, 6:81–92, 1991.
8. P. Murphy et M. Pazzani. Id2-of-3 : Constructive induction of m-of-n concepts for discriminators in decision trees. Technical Report 91-37, Department of Information and Computer Science - University of California at Irvine, 1991.
9. G. Pagallo et D. Haussler. Boolean feature discovery in empirical learning. *Machine Learning*, 5:71–99, 1990.
10. J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.
11. R. Rakotomalala. *Graphes d'Induction.* PhD thesis, University Claude Bernard - Lyon 1, December 1997.
12. L. Wehenkel. On uncertainty measures used for decision tree induction. In *Proceedings of Info. Proc. and Manag. Of Uncertainty*, pages 413–418, 1996.
13. A.P. White et W.Z. Liu. Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329, 1994.
14. D.A. Zighed, J.P. Auray et G. Duru. *Sipina : Methode et logiciel.* Lacassagne, 1992.