ATTRIBUTE SELECTION MEASURE IN DECISION TREE GROWING

Laviniu Aurelian Badulescu

University of Craiova, Faculty of Automation, Computers and Electronics,
Software Engineering Department

Abstract: One of the major tasks in Data Mining is classification. The growing of Decision Tree from data is a very efficient technique for learning classifiers. The selection of an attribute used to split the data set at each Decision Tree node is fundamental to properly classify objects; a good selection will improve the accuracy of the classification. In this paper, we study the behavior of the Decision Trees induced with 14 attribute selection measures over three data sets taken from UCI Machine Learning Repository.

Keywords: databases, decision trees, classification, error rates, machine learning.

1. INTRODUCTION

One of the major tasks in Data Mining (DM) is classification. The target of classification is to assign an object, from a data set, to a class, from a given set of classes, based on the attribute values of this object. The growing of Decision Tree (DT) from data is a very efficient technique for learning classifiers. The selection of an attribute used to split the data set at each node of the DT is fundamental to properly classify objects; a good selection will improve the accuracy of the classification. Different attribute selection measures were proposed in the literature (*e.g.* ID3 and C4.5 select the split attribute that minimizes the information entropy of the partitions, while SLIQ and SPRINT use the Gini index (Breiman *et al*. 1984; Wehenkel, 1996)) but it is not evident which of them will generate the best DT for a particular data set. In this paper, we study the behavior of the DT induced with 14 attribute selection measures over three data sets taken from UCI Machine Learning Repository.

The rest of the paper is organized as follows. First we present the attribute selection measures used in this paper and the related work in Section 2. Then we will discuss the behavior of the attribute selection measures on the growing, pruning and execution of the DT over the three databases: Abalone, Cylinder Bands and Image Segmentation (Section 3). Finally, we conclude the work with a discussion of future work (Section 4).

2. ATTRIBUTE SELECTION MEASURES. RELATED WORK

In this section we briefly describe the attribute selection measures used in our performance tests. For a data set S containing n records the information entropy

$$I(S) = -\sum_{i=1}^{K} P_i \cdot \log_2 P_i \qquad (1)$$

is defined as where P_i is the relative frequency of class i (there are K classes). For a split dividing S into m subsets: S_1 (n_1 records), … , S_m (n_m records) in accordance with the attribute test X, the information entropy is:

$$I_X(S_1,...,S_m) = -\sum_{i=1}^{m} \frac{n_i}{n} I(S_i)\,. \qquad (2)$$

The difference:

$$ing(X) = I(S) - I_X(S_1,...,S_m) \qquad (3)$$

measures the information that is gained by splitting S in accordance with the attribute test X. The attribute selection measure used by ID3 (Quinlan, 1986): gain criterion, selects an attribute test X to maximize *ing*, i.e., this attribute selection measure will choose an attribute X with the highest *ing*. *Ing* has one severe lack: a clear discrimination in favor of attribute tests with many outputs. For this reason C4.5 (Quinlan, 1993), instead of *ing*, uses another attribute selection measure: information gain ratio (*ingr*):

$$ingr(X) = \frac{ing(X)}{splitinfo(X)} \tag{4}$$

where

$$splitinfo(X) = -\sum_{i=1}^{m} \frac{n_i}{n} \cdot \log_2 \frac{n_i}{n} \tag{5}$$

represents the potential information generated by splitting data set S into m subsets S_i (Kantardzic, 2003).

The Gini index (*gini*) for a data set S is defined as

$$gini(S) = 1 - \sum_{i=1}^{K} P_i^2 \tag{6}$$

and for a split:

$$gini_X(S_1,...,S_m) = \sum_{i=1}^{m} \frac{n_i}{n} gini(S_i)\,. \tag{7}$$

Modified Gini index (*ginim*) (Kononenko, 1994) is highly correlated with gini index, so for an attribute X:

$$ginim_X(S_1,...,S_m) = \sum_{i=1}^{m} \left(\frac{\left(\frac{n_i}{n}\right)^2}{\sum_{i=1}^{m}\left(\frac{n_i}{n}\right)^2} \cdot \sum_{j=1}^{K} P_j^{\,2} \right) - \sum_{j=1}^{K} P_j^{\,2} \tag{8}$$

The Gini index may be normalized in order to remove a bias towards multi-valued attributes; it is obtained another attribute selection measure, symmetric Gini index (*ginis*) suggested in (Zhou and Dillon, 1991) and presented in a large context of many attribute selection measures by Borgelt (2000):

$$ginis = \frac{\sum_{i=1}^{K} P_i \sum_{j=1}^{m} P_j^2 + \sum_{i=1}^{m} \frac{n_i}{n} \sum_{i=1}^{K} P_i^2 - \sum_{i=1}^{K} P_i^2 - \sum_{j=1}^{m} \left(\frac{n_j}{n}\right)^2}{2 - \sum_{i=1}^{K} P_i^2 - \sum_{j=1}^{m} \left(\frac{n_j}{n}\right)^2}\,. \tag{9}$$

Testing many attribute selection measures (Wang *et al.* 2007) affirm that most of the measures yield reasonable results, however, the symmetric Gini index maximized the DT accuracy.

In accordance with Mantaras (1991) symmetric information gain ratio (*singr*) grows smaller DT than the *ingr*, particularly in the case of samples with multi-valued attributes. This attribute selection measure is established on a distance between partitions such that the selected attribute in a node induces the partition which is nearest to the proper partition of the subset of training samples matching to this node.

The same set S is divided in two partition: a partition S_A whose classes will be denoted A_i for $1 <= i <= n$ and a partition S_B, whose classes will be denoted B_j for $1 <= j <= m$. Let the probabilities:

$$P_i = P(A_i) \tag{10}$$

$$P_j = P(B_j) \tag{11}$$

$$P_{ij} = P(A_i \cap B_j) \tag{12}$$

$$P_{j/i} = P(B_j / A_i) \tag{13}$$

The average information of partition S_A (respectively S_B) which measures the randomness of the distribution of elements of S over the n (respectively m) classes of the partition is:

$$I(P_A) = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{14}$$

respectively,

$$I(P_B) = -\sum_{j=1}^{m} P_j \log_2 P_j \tag{15}$$

The mutual average information of the intersection of two partitions $P_A \cap P_B$ is:

$$I(P_A \cap P_B) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log_2 P_{ij} \tag{16}$$

and the conditional information**:**

$$I(P_A / P_B) = I(P_B \cap P_A) - I(P_A) =$$
$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log_2 \left(\frac{P_{ij}}{P_i}\right) = -\sum_{i=1}^{n} P_i \sum_{j=1}^{m} P_{j/i} \log_2 P_{j/i} \tag{17}$$

Finally it obtains two distances between partitions:

1. The metric distance measure:

$$d(P_A, P_B) = I(P_B / P_A) + I(P_A / P_B) \tag{18}$$

2. The normalization distance in [0,1]:

$$d_N(P_A, P_B) = \frac{d(P_A, P_B)}{I(P_A \cap P_B)} \tag{19}$$

Let K be the number of classes, A the number of attributes and V the number of values of a given attribute. Let $n_{..}$ the number of samples, $n_{i.}$ the number of samples from class C_i, $n_{.j}$ number of samples with the j^{th} value of the given attribute, and n_{ij} the number of samples from class C_i with the j^{th} value of the given attribute. Let the following probabilities: $P_{i\,j} = n_{ij} / n_{..}$, $P_{i.} = n_{i.} / n_{..}$, $P_{.j} = n_{.j} / n_{..}$, and $P_{i|j} = n_{ij} / n_{.j}$.

Let us consider the following entropies: I_K of the classes, I_A of the values of the given attribute, I_{KA} of the joint events class-attribute, and $I_{K|A}$ of the classes given the value of the attribute, where:

$$I_K = -\sum_{i} P_{i.} \log_2 P_{i.} \tag{20}$$

$$I_A = -\sum_j P_{.j} \log_2 P_{.j} \quad (21)$$

$$I_{KA} = -\sum_i \sum_j P_{ij} \log_2 P_{ij} \quad (22)$$

$$I_{K|A} = I_{KA} - I_A \,. \quad (23)$$

An attribute selection measure, the average absolute weight of evidence (*wevd*) (Kononenko, 1995; Michie, 1990), is based on plausibility which is an alternative to entropy. Let *odds = P/(1-P)*. For two-class problems the measure is defined as follows:

$$wevd_i = \sum_j P_{.j} \left| \log_2 \frac{odds_{i|j}}{odds_i} \right|, \quad i = 1, 2\,, \quad (24)$$

and it holds $wevd_1 = wevd_2$. For multi-class problems the measure is defined as follows:

$$wevd = \sum_i P_{i.}\ wevd_i \quad (25)$$

Attribute selection measure *relief* (Kira and Rendell, 1992; Kononenko, 1994) estimates the quality of attributes and deals well with strongly dependent attributes. This algorithm searches for the closest samples from the same class and the closest samples from different classes.

Baim (1988) has proposed the relevance (*rlv*) of an attribute, defined by:

$$rlv = 1 - \frac{1}{K-1} \sum_j \sum_{i \neq i_m(j)} \frac{n_{ij}}{n_{i.}} \quad (26)$$

where for a given attribute value *j*:

$$i_m(j) = arg \max_i \frac{n_{ij}}{n_i} \quad (27)$$

White and Liu (1994) have given the following formula for statistical measure χ^2 (*hi2*) with (*V*-1)(*K*-1) degrees of freedom:

$$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - n_{ij})}{E_{ij}}, \quad \text{where } E_{ij} = \frac{n_{.j}\ n_{i.}}{n_{..}} \quad (28)$$

K2 metric (Cooper and Herskovits, 1992) and BD metric (Buntine, 1991; Heckerman *et al.*, 1995) (*k2* & *bd*) were originally developed for learning Bayesian networks. Other attribute selection measures used in our experiments are stochastic complexity (*stc*) (Krichevsky and Trofimov, 1983; Rissanen, 1987) and symmetric specificity gain ratio (*ssgr*) (Borgelt and Kruse, 1997).

3. EXPERIMENTAL RESULTS

For the growing of the DT, several attribute selection measures have been tried. The experiments presume the growing of the DT on a training data set (in fact, there were induced 14 different DT using 14 attribute selection measures at the splitting of a DT node), the pruning of a DT (the 14 DT from the previous step are pruned, using two pruning methods: confidence level pruning and pessimistic pruning method) and finally, the DT execution on the test data set – different data of the ones used at the training of the DT - to calculate the classification error rate of each DT. Our tests use three well-known databases from (Newman et al. 1998): *Abalone*, *Cylinder Bands* and *Image Segmentation* from *Statlog Project*. Along with the performance of the file size for every DT induced with an attribute selection measure, we have also studied the behavior of the height and the number of nodes of every DT. The most important performance for the classification of the different DT, the classification accuracy on the test data, data completely unknown at the training of DT, has been noticed; this performance is expressed by classification error rate on the test data.

3.1 Abalone database

Number of instances: 4177 (3133 training, 1044 testing); number of attributes: 8 (continuous and nominal) and class *Rings* (with values: A, B, C); missing values: none.

From Table 1 we can see that the *stc* measure presents the smallest classification error rate (40.61%) and the *ginim* the biggest (44.83%). *Ingr* measure shows the biggest values for DT file size, height and number of DT nodes; after pruning, for this measure, the values for the file size and the number of nodes decrease, but the height of the DT remains the same (see Table 1, 2 and 3).

Table 1. Abalone. DT growing performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	67	36	980	42.62
gini	59	19	1002	44.16
ginim	53	15	997	44.83
ginis	57	18	985	41.95
hi2	64	21	1041	44.35
ing	64	24	1037	42.91
ingr	398	144	1402	44.54
k2	67	36	980	42.62
relief	53	20	951	43.10
rlv	61	14	1151	41.57
singr	78	29	1098	43.97
ssgr	77	34	1283	43.01
stc	100	58	1156	40.61
wevd	42	15	760	43.39

From Table 2 we can see that the *rlv* measure presents the smallest classification error rate (40.23%) and the *gini* the biggest (43.87%). It can be noticed a decrease of the classification error rates along with the pruning of DT for all the measures

(see Table 1, 2 and 3). The average of the classification error rates is 43.12% for unpruning DT, 42.19% for confidence level pruning DT and 42.48% for pessimistic pruning DT.

Table 2. Abalone. DT confidence level pruning performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	57	36	835	42.05
gini	45	18	767	43.87
ginim	42	14	817	42.34
ginis	48	18	834	41.86
hi2	50	21	823	43.39
ing	55	22	889	42.91
ingr	362	144	1202	43.30
k2	57	36	835	42.05
relief	43	17	788	41.19
rlv	48	14	901	40.23
singr	65	29	929	43.39
ssgr	45	15	788	41.00
stc	87	58	992	40.42
wevd	39	15	706	42.72

From Table 3 we can see that the *stc* measure presents the smallest classification error rate (40.33%) and the *ingr* the biggest (43.87%).

Table 3. Abalone. DT pessimistic pruning performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	62	36	910	42.43
gini	54	19	907	43.58
ginim	48	14	916	43.68
ginis	53	18	911	41.95
hi2	57	21	937	43.39
ing	62	24	997	42.72
ingr	379	144	1282	43.87
k2	62	36	910	42.43
relief	47	20	850	42.43
rlv	56	14	1042	40.42
singr	72	29	1025	43.49
ssgr	53	17	918	41.76
stc	95	58	1094	40.33
wevd	37	15	674	42.24

With four exceptions (*gini*, *ing*, *stc*, *wevd*) the confidence level pruning method produces DT with better classification error rates than pessimistic pruning method (see Table 2 and 3).

The correlation coefficients from Table 1, 2 and 3 between classification error rates and the file size (0.266, 0.240, 0.302), the height (0.137, 0.181, 0.219) and the number of nodes (0.006, 0.030, 0.035) of DT are very small indicating that file size, height or number of nodes of the DT have no influence upon classification accuracy.

3.2 Cylinder Bands database

Number of instances: 512 (412 training, 100 testing); number of attributes: 40 (20 numeric, 20 nominal) including the class attribute *band type* (with values: band, no band); missing values: in 302 samples.

Table 4. Cylinder Bands. DT growing performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	7	12	94	42
gini	12	2	357	81
ginim	12	2	357	81
ginis	12	2	357	81
hi2	12	2	357	81
ing	12	2	357	81
ingr	23	9	363	34
k2	7	12	94	42
relief	8	10	117	45
rlv	12	2	357	81
singr	12	2	357	81
ssgr	13	11	298	75
stc	7	14	95	44
wevd	6	9	70	55

From Table 4 we can see that the *ingr* measure presents the smallest classification error rate (34%).
From Table 4, 5 and 6 we can see that a group of 7 measure (*hi2*, *gini*, *ginim*, *ginis*, *ing*, *singr*, *rlv*) induce DT with very big classification error rates (81%) and aren't able to properly prune the DT having only one node and huge values (86%) for classification error rates.

From Table 5 we can see that two measures *k2* and *bd* achieve the smallest classification error rate (39%) an improved performance vs. unpruned DT (see Table 4).

From Table 6 we can see that two measures *k2* and *bd* achieve the smallest classification error rate (39%) an improved performance vs. unpruned DT (see Table 4).

Table 5. Cylinder Bands. DT confidence level pruning performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	7	12	84	39
gini	1	1	1	86
ginim	1	1	1	86
ginis	1	1	1	86
hi2	1	1	1	86
ing	1	1	1	86
ingr	2	8	11	80
k2	7	12	84	39
relief	6	10	94	50
rlv	1	1	1	86
singr	1	1	1	86
ssgr	3	10	25	83
stc	6	13	74	46
wevd	5	9	66	56

The correlation coefficients from Table 4, 5 and 6 between classification error rates and the file size (0.129, -0.985, -0.988), the height (-0.859, -0.867, -0.861) and the number of nodes (0.737, -0.974, -0.977) of DT are very big indicating that file size, height or number of nodes of the DT have a great influence upon classification accuracy. Indeed a DT with 1 node produces a huge classification error rate and this result is not in contradiction with the result of previous database.

Table 6. Cylinder Bands. DT pessimistic pruning performances

Measure	File size [kB]	Height	# nodes	Error rates [%]
bd	7	12	92	39
gini	1	1	1	86
ginim	1	1	1	86
ginis	1	1	1	86
hi2	1	1	1	86
ing	1	1	1	86
ingr	2	8	15	80
k2	7	12	92	39
relief	7	10	108	46
rlv	1	1	1	86
singr	1	1	1	86
ssgr	3	11	35	80
stc	6	13	82	46
wevd	5	9	66	56

3.3 Image Segmentation database

Number of instances: 6435 (4435 training, 2000 testing); number of attributes: 36 (all numeric) and the class attribute (with values: A, B, C, D, E and G); missing values: none.

For all the 14 measures the values of the performances (file sizes, heights, number of nodes of DT or classification error rates) are the same for the unpruned DT, for the confidence level pruned DT or for the pessimistic pruned DT (see Table 7). Confidence level pruning method improves the accuracy of the classification (15.95% vs. 16.80%), but the pessimistic pruning method keeps the same value for classification error rates like the unpruned DT (see Table 7). But with pessimistic pruning method we achieve better values for DT file size (103 kB vs. 107 kB) and better values for the DT nodes number (685 vs. 701). Confidence level pruning acquires also small values for the DT file size (70 kB vs. 107 kB), small values for the DT heights (62 vs. 81) and small values for the DT nodes number (529 vs. 701).

Table 7. Image Segmentation. DT performances

DT	File size [kB]	Height	# nodes	Error rates [%]
unpruned	107	81	701	16.80
confidence level pruned	70	62	529	15.95
pessimistic pruned	103	81	685	16.80

4. CONCLUSIONS AND FURTHER WORK

The experiments accomplished targeted the growing, the pruning, the execution of the unpruned and the pruned DT on the test data. We tried to study the behavior of DT grown with 14 different attribute selection measures and in the same time the classification accuracy on the test data of these trees. In our experiment we use 3 databases from literature with different types of attributes, numeric and nominal (*Abalone, Cylinder Bands*) and with missing values and large number of attributes (*Cylinder Bands*).

The best performance for the average classification error rate is accomplished by *k2, bd* (41.18%), *stc* (42.89%) and *relief* (44.62%) measures. They are followed by a group of two measures *wevd* (49.23%) and *ingr* (54.29%) with medium values for the classification error rates. The measures which have the worst performances for the average classification error rate on the test data are *ssgr* (60.63%), *rlv* (62.54%), *ginis* (63.13%), *ing* (63.59%), *ginim*

(63.98%), *singr* (63.98%), *hi2* (64.02%) and *gini* (64.10%).

The values of performance are influenced by the intrinsic features of each database; for certain databases some measures crash, for others all the measures have the same values for all the performances taken into account.

The experiments we wish to perform next will target much larger databases with many attributes and many samples on which we want to verify the performances of much more attribute selection measures.

ACKNOWLEDGMENTS

We want to note the assistance we received from Newman et al. (1998) and Ross D. King, Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, Scotland, for the Stalog databases that are a subset of the datasets used in the European Statlog Project.

REFERENCES

Baim, P. W. (1988), A method for attribute selection in inductive learning systems, *IEEE Trans. on PAMI*, **Volume 10**, **No. 6**, pp. 888-896.

Borgelt, C. (2000), *Data Mining with Graphical Models*, Ph. D. Thesis, Fakultat fur Informatik der Otto-von-Guericke-Universitat Magdeburg, p. 211, http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html.

Borgelt, C. and R. Kruse (1997), Evaluation Measures for Learning Probabilistic and Possibilistic Networks, *Proc. of the FUZZ-IEEE'97*, Barcelona, **Volume 2**, pp.669–676.

Breiman, L., J. Friedman, R. Olshen and C. Stone (1984), *Classification and Regression Trees*, Stanford University, Berkeley.

Buntine, W. (1991), Theory Refinement on Bayesian Networks, *Proc. 7th Conf. on Uncertainty in Artificial Intelligence (UAI 91)*, Morgan Kaufman, Los Angeles, pp. 52–60.

Cooper, G. F. and E. Herskovits (1992), A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning*, Springer, **Volume 9**, **No 4**, pp. 309–347.

Heckerman, D., D. Geiger and D. M. Chickering (1995), Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning,* Kluwer, Boston, **Volume 20**, **No. 3**, pp. 197–243.

Kantardzic, M. (2003), *Data Mining: Concepts, Models, Methods, and Algorithms*, Chapter 7.2, John Wiley & Sons, Louisville.

Kira, K. and L. Rendell (1992), A practical approach to feature selection, In: *Proc. Int. Conf. on Machine Learning*, D. Sleeman and P. Edwards (Ed), pp. 249-256, Morgan Kaufmann, Aberdeen.

Kononenko, I. (1994), Estimating Atributes: Analysis and extensions of RELIEF, In: *Proc. European Conf. on Machine Learning*, L. De Raedt and F. Bergadano (Ed), pp. 171-182, Springer Verlag, Catania.

Kononenko, I. (1995), On Biases in Estimating Multi-Valued Attributes, In: *Proc. of the 14th Int. Joint Conference on Artificial Intelligence (IJCAI'95),* C. S. Mellish (Ed.), pp. 1034–1040, Morgan Kaufmann, San Mateo, CA.

Krichevsky, R. E. and V. K. Trofimov (1983), The Performance of Universal Coding, *IEEE Trans. on Information Theory*, **Volume 27**, **No 2**, pp. 199–207.

Mantaras, R. L. de (1991), A Distance-based Attribute Selection Measure for Decision Tree Induction, *Machine Learning*, Kluwer, Boston, **Volume 6**, **No. 1**, pp. 81–92.

Michie, D. (1990), Personal Models of Rationality, *J. of Statistical Planning and Inference, Special Issue on Foundations and Philosophy of Probability and Statistics***, Volume 21**, pp. 381-399.

Newman, D.J., S. Hettich, C. L. Blake and C. J. Merz (1998), *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Depart. of Information and Computer Science.

Quinlan, J. R. (1986), Induction of Decision Trees, *Machine Learning*, Kluwer, Boston, **Volume 1**, pp. 81–106.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, Canada.

Rissanen, J. (1987), Stochastic Complexity, *J. of the Royal Statistical Society (Series B)*, **Volume 49**, **No. 3**, pp. 223-239.

Wang, X., D. D. Nauck, M. Spott and R. Kruse (2007), Intelligent data analysis with fuzzy decision trees, *Soft Computing*, **Volume 11**, **No. 5**, Springer-Verlag, pp. 439-457

Wehenkel, L. (1996), On Uncertainty Measures Used for Decision Tree Induction, *Proc. of the Int. Congress on Information Processing and Management of Uncertainty in Knowledge based Systems (IPMU96*), Granada, pp. 413-418.

White, A. P. and W. Z. Liu (1994), Bias in information-based measures in Decision Tree Induction, *Machine Learning*, Kluwer, Boston, **Volume 15**, pp. 321-329.

Zhou, X. and T. S. Dillon (1991), A statistical-heuristic Feature Selection Criterion for Decision Tree Induction, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, **Volume 13, No. 8,** pp. 834–841.