2018 22nd International Conference on System Theory, Control and Computing (ICSTCC)

An information entropy based splitting criterion better for the Data Mining Decision Tree algorithms

Bădulescu Laviniu Aurelian
Fac. of Automatics, Computer Science and Electronics
Univ. of Craiova
Craiova, Romania
badulescu@automation.ucv.ro

***Abstract*—This paper attempts to discover a better entropy-based splitting criterion for the induction of the Decision Trees (DT) than the entropy-based splitting criteria known so far. Eight entropy-based splitting criteria along with the performance tests carried out by three DT built on their basis, were taken into account: unpruned DT, pessimistically pruned DT and error-based pruned DT. All of these DT types were trained on seven databases, and then executed on their test data. Our experiments highlight the very good performances achieved by the square information gain ratio1 splitting criterion, which has shown always a better behavior than the information gain ratio criterion used in C4.5 algorithm and than other six information entropy based splitting criteria.**

***Keywords*—splitting criteria, Decision Trees, classification error rate, information entropy**

I. Background

For classification and prediction problems Decision Trees (DT) algorithms are some of the most efficient and intuitive Data Mining methods. One of the most important steps in developing an algorithm for inducing a DT is the choosing of splitting criterion. The information, and information-related entropy, provided the necessary support for the development of efficient splitting criteria, and classical examples are given by the splitting criteria used in DT algorithms ID3 and C4.5 (considered for a long time to be the most powerful Data Mining algorithm) developed by Ross Quinlan. Further developments, such as the *Mántaras distance*, have improved the performance of these entropy-based criteria.

This paper attempts to discover a better entropy-based splitting criterion than the entropy-based splitting criteria known so far. For this, eight entropy-based splitting criteria along with performance tests carried out by three DT built on their basis, were taken into account: unpruned DT, pessimistically pruned DT and error-based pruned DT. All of these DT types were trained on seven databases, and then executed on their test data, which were completely unknown on the moment of DT training.

The obtained results showed that although the *information gain* (*IG*) splitting criteria used in the ID3 algorithm and the *information gain ratio* (*IGR*) used in the C4.5 algorithm achieved very good results, they were systematically surpassed by a normalized variant of another entropy-based splitting criterion, *square information gain* (SIG), namely *square information gain ratio1* (*SIGR1*).

Next, in section Materials and Methods, we present the splitting criteria and the seven databases used in the experiments in this paper. In section Experimental Results and Discussions, we present and discuss the classification performance obtained from DT induction experiments, pruning them using two pruning methods (error-based and pessimistic) and performing them on the test data. At the end of the paper there is a Conclusion section.

II. Materials And Methods

Considering that $F = \{f_1, f_2, \ldots, f_i, \ldots, f_n\}$ denotes the collection of n input independent features, and L denotes the dependent feature or the class label, we can define a schema $R(F \cup L)$ on the relationship represented by the training dataset. A feature or attribute can be either categorical or continuous (numeric). Categorical or qualitative attributes are features whose values can be placed into distinct categories. For a categorical attribute f_i, $v(f_i) = \{a_{i,1}, a_{i,2}, \ldots, a_{i,|v(f_i)|}\}$ will denote its domain values. We will denote by $|v(f_i)|$ the number of values f_i can take. $v(L)=\{c_1, c_2, \ldots, c_{|v(L)|}\}$ will denote the values that the dependent attribute can take. Continuous features have an infinite number of values.

The Cartesian product $X = v(f_1) \times v(f_2) \times \ldots \times v(f_i) \times \ldots \times v(f_n)$ of all input attribute domains is called the *records space* and is the set of all possible records. We will consider a fragment of m records as the training dataset, with $x^m = \{x_1, x_2, \ldots, x_r, \ldots, x_m\}$ expresses the fraction of the records matching to the values of the independent features, and $L = L_1, L_2, \ldots, L_r, \ldots, L_m$ expresses the fraction of the m instances matching to the class labels.

An instance r is the pair $<x_r, L_r>$. $I(R) = (<x_1, L_1>, <x_2, L_2>, \ldots, <x_r, L_r>, \ldots, <x_m, L_m>)$ denotes the training dataset, where $x_r \in X$, $L_r \in v(L)$, $L_r = L_r(x_r)$, and r can take on values of 1 up to m. $\left|S_{y=c_k} I\right|$ denotes the number of records from I with the class label c_k, where $k = 1.. |v(L)|$, $|I|$ is the total number of instances from I, and $\left|S_{L=c_k} I\right| / |I|$ is the relative frequency of records with the class label c_k in the I training dataset. Then the entropy of the I set is:

$$E(L,I) = -\sum_{c_k \in v(L)} \frac{\left|S_{L=c_k} I\right|}{|I|} \cdot \log_2 \frac{\left|S_{L=c_k} I\right|}{|I|} \qquad (1)$$

978-1-5386-4444-7/18/$31.00 ©2018 IEEE

A decision tree is a tree-like structure in which root-node represents entire dataset, each internal node is splitted into two or more sub-nodes by a test on a feature, each edge represents the result of the test, and each leaf or terminal node represents a class label. If we choose the attribute f_i to split a node, then $|v(f_i)|$ sub-nodes result, each one having $S_{f_i=a_{i,j}} I$ instances from I, i.e., those instances of I that have the feature f_i with the value $a_{i,j}$, where j takes values from 1 to $|v(f_i)|$. The partition entropy $S_{f_i=a_{i,j}} I$ of I is:

$$E(L,S_{f_i=a_{i,j}} I) = -\sum_{c_k\in v(L)} \left(\frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}} I\right|}{|I|} \cdot \log_2 \frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}} I\right|}{|I|} \right) \tag{2}$$

If $\left|S_{f_i=a_{i,j}} I\right| / |I|$ denotes the relative frequency of the instances with value $a_{i,j}$ for the feature f_i, then

$$E_d = -\sum_{a_{i,j}\in d(f_i)} \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \cdot E(L,S_{f_i=a_{i,j}} I) \tag{3}$$

defines the entropy of the splitting of I in $|v(f_i)|$ partitions, at the division of a node on the f_i feature base. The difference:

$$\begin{aligned} &IG(L,S_{f_i=a_{i,j}} I) = \\ &= E(L,I) - \sum_{a_{i,j}\in v(f_i)} \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \cdot E(L,S_{f_i=a_{i,j}} I) = \\ &= E(L,I) + E(f_i,I) - E(L,S_{f_i=a_{i,j}} I) \end{aligned} \tag{4}$$

expresses the *information gain* (*IG*) splitting criterion, as [1], and represents the information that is reached by splitting I on the basis of the attribute f_i. In (4) $E(fi, I)$, with following expression:

$$E(f_i,I) = -\sum_{a_{i,j}\in v(f_i)} \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \cdot \log_2 \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \tag{5}$$

defines the potential information created by the segmentation of the dataset I based on the feature f_i in $|v(f_i)|$ subsets $S_{f_i=a_{i,j}} I$, where j takes values from *1* to $|v(f_i)|$. The *IG* splitting criterion was used by Quinlan to implement the ID3 algorithm.

The criteria based on information entropy that we are discussing below are attempts to refine the *IG* criterion that prefers multi-valued attributes. *Balanced information gain* (*BIG*), as in [2], is an endeavor to balance *IG*'s tendency to favor splittings with many paths:

$$\begin{aligned} &BIG(L,S_{f_i=a_{i,j}} I) = \\ &= \frac{E(L,I) + E(f_i,I) - E(L,S_{f_i=a_{i,j}} I)}{\log_2 |v(f_i)|} \end{aligned} \tag{6}$$

We can notice that for binary divisions (i.e., $|v(fi)| = 2$), the *BIG* criterion coincides with the *IG* criterion, because, in this case, the denominator $\log_2|v(fi)|$ is 1. By introducing *quadratic entropy*, as in [3], as:

$$E^2(L,I) = 2 - 2\sum_{c_k\in v(L)} \left(\frac{\left|S_{L=c_k} I\right|}{|I|} \right)^2 \tag{7}$$

and using the *IG* model of (4), we define another splitting criterion, namely, *square information gain* (*SIG*):

$$\begin{aligned} &SIG(L,S_{f_i=a_{i,j}} I) = E^2(L,I) - \\ &- \left(\sum_{a_{i,j}\in v(f_i)} \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \cdot E(L,S_{f_i=a_{i,j}} I) \right)^2 \end{aligned} \tag{8}$$

or

$$\begin{aligned} &SIG(L,S_{f_i=a_{i,j}} I) = \\ &= E^2(L,I) + E^2(f_i,I) - E^2(L,S_{f_i=a_{i,j}} I) \end{aligned} \tag{9}$$

This criterion, similar to *IG*, undergoes from the same problems as this one, that is, it will be biased towards attributes with many values, which it will prefer in the detriment of those with few values. Due to the squaring, the negative behavior of the *SIG* criterion worsens in comparison to the *IG* criterion behavior. Therefore, following the model of balancing the *IG* criterion by the *BIG* criterion and for the *SIG* criterion, we will define a balanced variant called the *balanced square information gain* (*BSIG*). This criterion is an attempt to balance the tendency of *SIG* criterion to excessively favor the multiple paths splittings. The *BSIG* criterion has the following formula:

$$\begin{aligned} &BSIG(L,S_{f_i=a_{i,j}} I) = \\ &= \frac{E^2(L,I) - \left(\sum_{a_{i,j}\in v(f_i)} \frac{\left|S_{f_i=a_{i,j}} I\right|}{|I|} \cdot E(L,S_{f_i=a_{i,j}} I) \right)^2}{\left(\log_2 |v(f_i)|\right)^2} \end{aligned} \tag{10}$$

or

$$BSIG(L,S_{f_i=a_{i,j}}I)= \\ =\frac{E^2(L,I)+E^2(f_i,I)-E^2(L,S_{f_i=a_{i,j}}I)}{\left(\log_2|v(f_i)|\right)^2} \tag{11}$$

We note that for the binary splitting (i.e., $|v(fi)| = 2$), the *BSIG* criterion coincides with the *SIG* criterion, because, in this case, the denominator $(\log_2|v(fi)|)^2$, has the value 1. Besides balancing, we can also eliminate the *SIG* criterion preference for multi-valued attributes by normalization. The normalization of the *SIG* criterion gives us two splitting criteria: *square information gain ratio 1* (*SIGR1*) and *square information gain ratio 2* (*SIGR2*). The formula of the *SIGR1* criterion is as follows:

$$SIGR1(L,S_{f_i=a_{i,j}}I)= \\ =\frac{E^2(L,I)+E^2(f_i,I)-E^2(L,S_{f_i=a_{i,j}}I)}{E^2(f_i,I)} \tag{12}$$

The formula of the *SIGR2* criterion is as follows:

$$SIGR2(L,S_{f_i=a_{i,j}}I)= \\ =\frac{E^2(L,I)+E^2(f_i,I)-E^2(L,S_{f_i=a_{i,j}}I)}{E^2(L,I)+E^2(f_i,I)} \tag{13}$$

To normalize the *IG* criterion, Quinlan proposed for C4.5 the *information gain ratio* (*IGR*), as [1], that normalizes *IG* as follows:

$$IGR(L,S_{f_i=a_{i,j}}I)= \\ =\frac{E(L,I)+E(f_i,I)-E(L,S_{f_i=a_{i,j}}I)}{E(f_i,I)} \tag{14}$$

or

$$IGR(L,S_{f_i=a_{i,j}}I)=\frac{IG(L,S_{f_i=a_{i,j}}I)}{E(f_i,I)} \tag{15}$$

where

$$E(f_i,I)=-\sum_{a_{i,j}\in v(f_i)}\frac{\left|S_{f_i=a_{i,j}}I\right|}{|I|}\cdot\log_2\frac{\left|S_{f_i=a_{i,j}}I\right|}{|I|} \tag{16}$$

is the potential information supplied by the splitting of the dataset I on the basis of the feature f_i, in $|v(f_i)|$ subsets under the form $S_{f_i=a_{i,j}}I$, where j takes values from 1 to $|v(f_i)|$. We can notice that the *IGR* criterion is not defined if the denominator is zero. Also, the *IGR* criterion tends to favor the attributes for which the denominator is very small. Quinlan showed that the *IGR* criterion tends to behave better than the *IG* criterion, both in terms of precision and of the aspects of the classifier's complexity, as [1].

Reference [4] shows the *Mántaras distance*, a splitting criterion that constructs lower DT than the *IGR* criterion proposed by Quinlan for C4.5. A better performance of the proposed *Mántaras* criterion is particularly noticeable for records that contain attributes with many values. Using the *Mántaras distance* we define the *Mántaras information gain ratio* (*MIGR*) splitting criterion:

$$MIGR(L,S_{f_i=a_{i,j}}I)=\frac{E(L,I)+E(f_i,I)-E(L,S_{f_i=a_{i,j}}I)}{E(L,S_{f_i=a_{i,j}}I)}= \\ =\frac{\sum_{c_k\in v(L)}\frac{\left|S_{L=c_k}I\right|}{|I|}\cdot\log_2\frac{\left|S_{L=c_k}I\right|}{|I|}}{\sum_{c_k\in v(L)}\frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}}I\right|}{|I|}\cdot\log_2\frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}}I\right|}{|I|}}+ \\ +\frac{\sum_{a_{i,j}\in v(f_i)}\frac{\left|S_{f_i=a_{i,j}}I\right|}{|I|}\cdot\log_2\frac{\left|S_{f_i=a_{i,j}}I\right|}{|I|}}{\sum_{c_k\in v(L)}\frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}}I\right|}{|I|}\cdot\log_2\frac{\left|S_{L=c_k\ AND\ f_i=a_{i,j}}I\right|}{|I|}}-1 \tag{17}$$

The experiments presented in this paper continue the research on the splitting criteria previously made on two databases: *Census-Income (KDD)* [5] and *Forest Covertype* [6]. The experiments in this paper focus on the following seven databases:

Abalone database [7]: number of cases: 4177 (train = 3133, test = 1044); number of attributes: 8 (continuous and categorical), and the class attribute with three values; missing values: none.

Cylinder Bands database [7]: number of cases: 512 (train = 412, test = 100); number of attributes: 40 (continuous and categorical), and the class attribute with two values; missing values: in 302 cases.

Landsat Satellite database from *Statlog Project* [7]: number of cases: 6435 (train = 4435, test = 2000); number of attributes: 36 (all continuous), and the class attribute with six values; missing values: none.

Monk's Problem database (only *Monk-1 problem*) [7]: number of cases: 124 for training and 432 for testing; number of attributes: 6 (continuous), and the class attribute with two values; missing values: none.

Adult database [7]: number of cases: 48,842 (train = 32,561, test = 16,281); number of attributes: 15, and the class attribute with two values; missing values are confined to three attributes. There are 6 duplicates or conflicting cases.

Census-Income (KDD) [7]: number of cases: 299,285 (train = 199,523, out of which duplicated or contradictory cases: 46,716; test = 99,762, out of which duplicated or contradictory cases: 20,936); number of attributes: 40 (continuous and categorical), and the class attribute with two values.

Forest Covertype database is in UCI KDD Archive (http://kdd.ics.uci.edu, copyright J. A. Blackard & Colorado State University): number of cases: 581,012 (train =15,120, test = 565,892); number of attributes 54 (continuous and categorical) and the class attribute with seven values; missing values: none.

III. Experimental Results And Discussions

We conducted experiments on seven databases, each building three types of DT: unpruned, error-based pruned, or pessimistically pruned; each DT was tested on the test data. Each of the three types of DT was induced in eight variants using eight splitting criteria. For the performance test part, we chose to highlight a key point of algorithms based on DT, namely the splitting criteria that determine the attribute selection for establishing the test that decides the ramification of a node, trying to find the criterion that consistently proves the best behavior in classification and prediction. The tests revealed a diversity of performance. On some datasets certain criteria have proven their efficiency, while others have been more uniform in terms of the splitting criteria, but the pruning methods have been those that have shown different performance. Table I lists all values of the classification error rate on the test data for all the eight splitting criteria applied to seven databases, using three categories of DT: unpruned DT, pessimistically pruned DT, and error-based pruned DT. In Table I, the second to last line corresponds to the average of the classification error rates for each of the 8 splitting criteria, and the last line contains the standard deviation of the classification error rate on the test data.

A. Average

The values of the average classification error rates on the test data is presented on the second to last line in Table 1. It has been found that the DT induced with the *SIGR1* criterion best classifies, with an average classification error rate on the test data of 18.59%. The DT performances with the other criteria are placed at a distance, having values close to each other for the average of the classification error rate. The weakest average performance for the classification error rate is achieved by the DT built with the *BSIG* splitting criterion, having an average value of 31.12% for the classification error rate.

B. Standard deviation

The conclusions provided by arithmetic mean values are often not eloquent, because the arithmetic mean can smooth out extreme values, while the standard deviation indicates the way these values are grouped together. From the last line of Table I, we find that the *SIGR1* criterion also has the smallest value for the standard deviation (14.11), so the lowest scattering of the classification error rate values on the test data around their average. This result indicates that the performance of the DT induced with the *SIGR1* criterion is not affected too much by the database features. Table I shows that the other DT have higher values for the standard deviation of the classification error rate than the DT induced with the *SIGR1* criterion. Note that the next criterion, which simultaneously obtains the second value at the average of the classification error rate (27.36%) and the second value at the standard deviation (21.94%), is the *IGR* (associated with the C4.5 algorithm).

C. win/tie/loss

In addition to the arithmetic mean, many authors (e.g., [8]) prefer the *win/tie/loss* method to compare the classification performance of the various algorithms used in experiments on different databases. This technique is based on counting cases where a classification model has achieved a better (*win*), equal (*tie*) or less (*loss*) performance than another model considered as reference. So we used 8 times Table I and every time we considered as reference another criterion that all the other criteria were reported using the

TABLE I. Classification Error Rate On Test Data For The 3 Types Of Dt, 7 Databases, And 8 Splitting Criteria [%]

Database	Type of pruning	Splitting criterion							
		SIGR1	*IGR*	*BIG*	*IG*	*MIGR*	*SIGR2*	*SIG*	*BSIG*
Abalone	unpruned	40.13	44.54	43.20	42.91	43.97	43.01	40.71	40.71
	err-based	39.56	43.30	43.01	42.91	43.39	42.82	40.23	40.23
	pess	39.94	43.87	42.91	42.72	43.49	42.72	40.23	40.23
Cylinder Bands	unpruned	16.00	34.00	81.00	81.00	81.00	81.00	81.00	81.00
	err-based	15.00	80.00	86.00	86.00	86.00	86.00	86.00	86.00
	pess	15.00	80.00	86.00	86.00	86.00	86.00	86.00	86.00
Image Segmentation	unpruned	16.80	16.80	16.80	16.80	16.80	16.80	16.80	16.80
	err-based	15.95	15.95	15.95	15.95	15.95	15.95	15.95	15.95
	pess	16.80	16.80	16.80	16.80	16.80	16.80	16.80	16.80
Monk's Problem	unpruned	0.00	11.11	10.65	10.65	10.65	11.57	12.50	12.50
	err-based	0.00	11.11	13.89	13.89	13.89	10.19	12.50	12.50
	pess	0.00	11.11	10.65	10.65	10.65	10.19	12.50	12.50
Adult	unpruned	18.10	16.87	17.63	18.01	16.89	19.24	19.97	19.97
	err-based	14.12	13.41	13.80	15.04	13.91	16.25	17.36	17.31
	pess	16.69	15.52	16.70	16.88	16.03	18.21	18.94	18.94
Census Income	unpruned	5.84	5.89	6.30	6.34	5.86	6.21	5.57	6.75
	err-based	4.84	4.81	4.66	5.14	4.72	5.31	5.57	5.57
	pess	5.33	5.06	5.76	5.57	5.15	5.73	6.03	6.03
Forest Covertype	unpruned	36.91	34.91	33.17	33.17	33.97	36.71	39.36	39.36
	err-based	36.74	34.57	32.60	32.60	33.65	36.37	39.03	39.03
	pess	36.59	34.83	32.90	32.90	33.88	36.59	39.42	39.42
average		18.59	27.36	30.02	30.09	30.13	30.65	31.07	31.12
standard deviation		14.11	21.94	25.83	25.75	25.97	25.85	25.54	25.48

win/tie/loss technique. For space reasons, I have not presented all the 8 tables, but only Table II that shows only the *win/tie/loss* column from each of the 8 tables.

When Table II is read on lines, for example, on the *IG* criterion line, alongside the *SIGR1* column, we find the *win/tie/loss* values as 4/3/14. This means that the *IG* criterion performs better than the *SIGR1* criterion for the accuracy of classification on test data in 4 cases, and has an equal performance in 3 cases and a poor performance in 14 cases.

When Table II is read on columns, for example, if we want to compare the *BIG* criterion with the *IG* criterion, we go to the *BIG* criterion column on the line corresponding to the *IG* criterion and read 4/12/5. These values tell us that the performance of the *BIG* criterion is better than the performance of the *IG* criterion in 5 cases, equal in 12 and weaker in 4.

In Table III we summed up the *win/tie/loss* values from Table II, for each criterion, and presented them sorted in three ways. At first, descending based on values of *win*, that is, the number of cases in which that criterion was better than the others. Then ascending, according to the value of the *loss*, i.e., the number of cases in which that criterion had a lower performance than the other criteria. It is obvious that the criterion that had the least cases in which it behaved less well than the others should be higher. The last sorting of the *win/tie/loss* values shown in Table III is a descending sorting according to the value of the difference: *win-loss*.

We consider that it is important for a criterion to have not only a high *win* value but also a high value of the difference: *win-loss*. This difference orders the criteria by placing higher than those that had a high value for *win*, but at the same time, a low value for *loss*. It can be seen that in any of the three types of ordering in Table III, the *SIGR1* criterion is always in the first position. This leads to two conclusions: the *SIGR1* criterion has the highest number of cases when it has a better performance value than all the other criteria, and the *SIGR1* criterion presents the smallest number of cases when it has a lower performance value than the other criteria.

Symmetrically to the *SIGR1* criterion, which occupies the first place irrespective of the criterion chosen for ordering, the *BSIG* criterion always occupies the last place in Table III.

TABLE II. THE WIN/TIE/LOSS VALUES FOR THE CLASSIFICATION PRECISION, CONSIDERING EACH OF THE 8 CRITERIA A BENCHMARK COMPARISON FOR THE OTHER 7

Criterion	*BIG*	*SIG*	*BSIG*	*IGR*	*SIGR1*	*SIGR2*	*MIGR*
IG	4/12/5	10/6/5	11/6/4	8/3/10	4/3/14	10/7/4	6/9/6
	BIG	10/6/5	11/6/4	9/3/9	6/3/12	8/6/7	8/9/4
		SIG	1/19/1	4/3/14	1/3/17	4/6/11	5/6/10
			BSIG	3/3/15	0/3/18	3/6/12	4/6/11
				IGR	8/3/10	13/3/5	9/3/9
					SIGR1	15/4/2	10/3/8
						SIGR2	5/6/10

TABLE III. GLOBAL VALUES FOR THE WIN/TIE/LOSS AT THE ACCURACY OF THE CLASSIFICATION, SORTED BY DIFFERENT CRITERIA

#	Cri-terion	win/tie/loss	Cri-terion	win/tie/loss	Cri-terion	win/tie/loss
1	*SIGR1*	96/22/29	*SIGR1*	96/22/29	*SIGR1*	96/22/29
2	*IGR*	78/21/48	*BIG*	57/45/45	*IGR*	78/21/48
3	*MIGR*	58/42/47	*MIGR*	58/42/47	*BIG*	57/45/45
4	*BIG*	57/45/45	*IGR*	78/21/48	*MIGR*	58/42/47
5	*IG*	53/46/48	*IG*	53/46/48	*IG*	53/46/48
6	*SIGR2*	46/38/63	*SIGR2*	46/38/63	*SIGR2*	46/38/63
7	*SIG*	25/49/73	*SIG*	25/49/73	*SIG*	25/49/73
8	*BSIG*	19/49/79	*BSIG*	19/49/79	*BSIG*	19/49/79
Descending sort after win value			***Ascending sort after loss value***		***Descending sort after win-loss difference value***	

D. Geometric mean

As a method of comparing the relative performance of two algorithms, some authors (e.g., [9]) propose the geometric mean of the ratio of the classification error rate for the two algorithms considered on several databases.

The geometric mean of a set of ratios a_1/b_1, a_2/b_2, … , a_n/b_n has the property that if it is higher than 1 (i.e., the values a_1, a_2, … , a_n correspond to a more efficient algorithm than the algorithm that obtains the values b_1, b_2, … , b_n), then the geometric mean of the ratios b_1/a_1, b_2/a_2, …, b_n/a_n is lower than 1 and vice versa. We considered that a_1, a_2, … , a_n represents the values of the classification error rate on the test data for an DT induced with a certain criterion on *n* databases and b_1, b_2, …, b_n represents the values of the classification error rate on the test data for an DT induced by another criterion on the same *n* databases. We have processed 8 times the data in Table I by calculating the geometric mean of the ratio of the error classification rate, considering, in turn, as a reference, each criterion, that we compared with all the other 7 criteria. Due to the fact that this statistical indicator cannot be calculated if the classification error rate is 0, we have replaced in Table I the three occurrences of the value 0 in the *Monk's Problem* database for the *SIGR1* criterion with a small value: 0.00001. After processing we obtained, for each of the eight criteria, two numerical values: the first represents the number of times the criterion considered as a reference had better values than the other criteria (*win* value), and the second value represents the number of times the criterion considered as a reference had values less good than the other criteria (*loss* value).There are no cases of equality. These two sequences of *win & loss* values are shown in Table IV. Table IV shows

TABLE IV. COMPARISON OF CLASSIFICATION ERROR RATE ON TEST DATA BASED ON GEOMETRIC MEAN

Cri-terion	*SIGR1*	*IGR*	*MIGR*	*BIG*	*IG*	*SIGR2*	*SIG*	*BSIG*
win	7	6	5	4	3	2	1	0
loss	0	1	2	3	4	5	6	7

that the performances of DT built with the *SIGR1* criterion surpass the performances of the DT built with the other 7 criteria both in the *win* column and in the *loss* column, in all the 21 cases considered in the experiments.

IV. Conclusions

The information entropy splitting criteria have been developed while trying to improve the *IG* criterion proposed for the ID3 algorithm. Our article presented comparatively the performances of classifying eight such criteria.

Generally, it is considered that the best splitting criterion based on the information entropy is the *IGR*. Out of the eight criteria used in the classification over seven databases, our tests have highlighted another criterion, *SIGR1*, which systematically shows better classification performance than the *IGR* criterion.

The *SIG* criterion favors many output attributes for the test, a problem inherited from the *IG* criterion on which it is based. The problem is aggravated by *SIG* due to the squaring. This situation is noticed in the results presented in Section III where the *SIG* criterion has less good performances than the *IG* criterion.

The attempt to solve this problem led to the development of two normalized criteria: the *SIGR1* and *SIGR2* criteria. Performance tests have shown that only one of the two normalization variants obtains spectacular results: *SIGR1*. This criterion, as shown by the values obtained in Section III, has systematically the best classification behavior, always better than the behavior of the *IGR* criterion, which is always behind it, in second place.

The other normalized version of the *SIG* criterion, the *SIGR2* criterion, performs better than the *SIG* criterion, proving that criterion normalization has improved the performances, but the improvement is not as obvious as in the *SIGR1* criterion. The *SIGR2* criterion obtains modest performances, placing itself systematically on the 6th place among the 8 criteria considered.

Due to the fact that the *IG* criterion favors the test with many outputs, it has been normalized by Quinlan who proposed, for the C4.5 algorithm, the *IGR* splitting criterion that normalizes *IG*. From the values presented in Section III we can observe the very good behavior of the *IGR* criterion and the fact that it is almost always present immediately after the *SIGR1* criterion.

The *MIGR* normalized criterion eliminates the preference of the *IG* criterion for multi-valued attributes. This fact is also demonstrated experimentally through the values that were presented in Section III. Thus, it can be observed that the *MIGR* criterion usually achieves better performances than the *IG* criterion. Instead, the *MIGR* criterion performs less well than the *IGR* criterion.

Sometimes the performances of the *MIGR* criterion are surpassed by the *BIG* criterion. We see a better systematic position in the rankings from Section III of the *BIG* criterion against the *IG* criterion. *BIG* criterion has been built so as to balance the tendency of the *IG* criterion to favor multi-path splitting. The values from the experiments show that the *IG* criterion has been improved by the *BIG* criterion.

The balanced variant of the *SIG* criterion, the *BSIG* criterion, has failed to balance the *SIG* criterion's tendency to excessively favor many paths splittings. Thus, the *BSIG* criterion obtains the weakest performances in the experiments presented.

Our experiments highlight the very good performances achieved by the *SIGR1* splitting criterion, which has always a better behavior than the *IGR* criterion. After studying DT classification performance with the eight splitting criteria, for three types of DT, on seven databases, we find that, regardless of the method of comparison considered, the *SIGR1* criterion induces the DT best performers.

Our researches will continue with the testing of other splitting criteria on other datasets to confirm or invalidate the findings of this paper.

Acknowledgment

We want to note the assistance we received from [7] and Ross D. King, Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, Scotland.

References

[1] L. Rokach, and O. Maimon, Data mining with decision trees. Theory and applications, Series in Machine Perception and Artificial Intelligence - vol. 81, World Scientific Pub. , New Jersey, 2015.

[2] S. B. Kim and D. L. Gillen, “A Bayesian adaptive dose-finding algorithm for balancing individual- and population-level ethics in Phase I clinical trials”, in Sequential Analysis, vol. 35, issue 4, N. Mukhopadhyay, Eds., Philadelphia:Taylor & Francis, 2016, pp. 423-439.

[3] W. M. Czarnecki and J. Tabor, “Extreme entropy machines: robust information theoretic classification”, in Journal Pattern Analysis & Applications, vol. 20, issue 2, S. Singh, Eds., London: Springer-Verlag, pp. 383-400, May 2017.

[4] E. Armengol, À. García-Cerdaña, P. Dellunde, “Experiences using decision trees for knowledge discovery”, in Fuzzy Sets, Rough Sets, Multisets and Clustering, V. Torra, A. Dahlbom and Y. Narukawa, Eds., SCI, volume 671, Springer, 2017, p.174.

[5] L. A. Bădulescu, “Pruning methods and splitting criteria for optimal decision trees algorithms”, in Annals of the Univ. of Craiova, vol. 13 (40), no. 1, pp. 15-19, 2016.

[6] L. A. Bădulescu, “Data mining classification experiments with decision trees over the Forest Covertype database”, in Proc. of 21st Int. Conf. on System Theory, Control and Computing (ICSTCC 2017), IEEE Conferences, Sinaia, Romania, pp. 236–241, 2017.

[7] D. Dua and E. Karra Taniskidou. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017.

[8] A. Narassiguin, H. Elghazel and A. Aussem, “Dynamic ensemble selection with probabilistic classifier chains”, in Machine Learning and Knowledge Discovery in Databases, M. Ceci, J. Hollmén, L. Todorovski, C. Vens and S. Džeroski, Eds. ECML PKDD 2017. Lecture Notes in Computer Science, vol 10534. Springer, Cham, 2017, pp. 169–186.

[9] O. A. Alzubi, J. A. Alzubi, S. Tedmori, H. Rashaideh and O. Almomani, “Consensus-based combining method for classifier ensembles”, in Int. Arab J. Inf. Technol. 15(1), 2018, pp. 76-86.