A New Criterion in Selection and Discretization of Attributes for the Generation of Decision Trees

Byung Hwan Jun, *Member, IEEE Computer Society*, Chang Soo Kim, Hong-Yeop Song, *Member, IEEE*, and Jaihie Kim, *Member, IEEE Computer Society*

Abstract—It is important to use a better criterion in selection and discretization of attributes for the generation of decision trees to construct a better classifier in the area of pattern recognition in order to intelligently access huge amount of data efficiently. Two well-known criteria are gain and gain ratio, both based on the entropy of partitions. We propose in this paper a new criterion based also on entropy, and use both theoretical analysis and computer simulation to demonstrate that it works better than gain or gain ratio in a wide variety of situations. We use the usual entropy calculation where the base of the logarithm is not two but the number of successors to the node. Our theoretical analysis leads some specific situations in which the new criterion works always better than gain or gain ratio, and the simulation result may implicitly cover all the other situations not covered by the analysis.

Index Terms—Decision-tree generators, attribute selection, discretization, grouping, gain, gain ratio, normalized gain, entropy.

1 Introduction

THE generation of a decision tree has been used as a method of machine learning for efficient acquisition of knowledge from mass amounts of data. Therefore, in the beginning, most of the research were focused on the study of how to select an attribute among many possibilities which have symbolic, nominal, and/or categorical values. In this paper, we propose a new criterion based on entropy which would overcome the limitation of such well-known criteria as gain or gain ratio. Using both theoretical analysis and computer simulation, we demonstrate its efficiency in the discretization and selection of attributes for the decision-tree generator whose partitioning approach is less restricted.

Quinlan used the difference of the entropies before and after a partition as a criterion for attribute selection in ID3 [1] and called it gain. This criterion has a tendency to prefer the attribute whose partition is more refined. It not only increases the size of the decision tree but also increases the error rate for unseen samples [1]. As an alternative, Quinlan modified gain in order to overcome these problems, and proposed gain ratio which is defined as the gain calibrated by some measure called IV. The generator that gain ratio is substituted for gain in ID3 was called ID3-IV [1]. These days, some of such well-known decision-tree generators as GID3 [2], GID3* [3], C4 [4], and C4.5 [5] have adopted gain ratio as a criterion for attribute selection. On the other hand, in the area of pattern recognition, the use of continuous attributes becomes more and more frequent, and causes a new problem of discretizing them. Initially, the studies of better criteria in the discretization and selection of attributes were performed independently, but later they were joined and studied simultaneously. It seems to be more appropriate to do this together because the selection of an attribute is in fact a selection of one of differently discretized (or grouped) attributes. These days, there are wide varieties of the above schemes, and C4.5 is known to be the most frequently referred decision-tree generator.

This paper is organized as follows. Section 2 reviews gain and gain ratio, as defined by Quinlan, and proposes a new criterion, called "normalized gain" based also on entropy. Some motivations and theoretical analysis to justify this new definition are also in this section. Section 3 describes the process of simulation in detail and presents its result in a table. Finally, we conclude in Section 4.

2 Old and New Criteria for Discretization and Selection of Attributes

In the generation of a decision tree for a classification in a top-down nonbacktracking style, an important point is to use a better criterion in the selection of an attribute among many possibilities. Here, the attribute selection should be interpreted in a broad sense including discretization. In this section, the entropy of a partitioned sample set will be briefly discussed, two criteria based on entropy will be reviewed, and, then, a new criterion will be proposed.

2.1 Entropy of a Partition

Let a sample set S be composed of k classes $c_1, c_2, ..., c_k$, having probabilities $p_1, p_2, ..., p_k$, respectively. Then, the entropy of S is defined as

$$\mathrm{Ent}(S) \triangleq -\sum_{i=1}^{k} p_i \log_2 p_i \tag{1}$$

Here, the unit of the above quantity is "bit" since the number two is used as the log base. Entropy in (1) is known to be a unique function which satisfies the four axioms of uncertainty [6] and represents the average amount of information when coding each class into a codeword with ideal length according to its probability [7].

In a classification problem, we need to evaluate the entropy of a given set partitioned by the selected attribute. Let an attribute A divide S into n disjoint subsets $S_1, S_2, ..., S_n$. Then, the entropy $E(A, S)$ of S partitioned by A is defined as the weighted average of entropies of subsets S_i for $i = 1, 2, ..., n$. That is, a subset S_i now has the entropy of partition $-\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$, where $p_{ij} \triangleq \frac{|S_i \cap c_j|}{|S_i|}$, $j = 1, 2, ..., k$. Here, p_{ij} is the size of samples of class c_j in S_i relative to the size of S_i. The weight is the relative size of S_i to S, and this gives

$$E(A, S) \triangleq -\sum_{i=1}^{n} P_i \sum_{j=1}^{k} p_{ij} \log_2 p_{ij} \tag{2}$$

where

$$P_i \triangleq \frac{|S_i|}{|S|}, \quad i = 1, 2, \ldots, n.$$

Two important properties of entropy are as follows:

1) if the number of classes is fixed, entropy increases as the probability distribution of classes becomes more uniform and
2) if the probability distribution of classes is uniform, entropy increases logarithmically as the number of classes in a sample set increases.

- *B.H. Jun is with the Department of Computer Science, Kongju National University, 182 Shinkwan-Dong, Kongju City, Chungnam, 314-701, Korea. E-mail: bhjun@kcs.kongju.ac.kr.*
- *C.S. Kim, H.-Y. Song, and J. Kim are with the Department of Electronic Engineering, Yonsei University, 134 Shinchon-Dong, Sudaemoon-Ku, Seoul, 120-749, Korea. E-mail: hammer@seraph.yonsei.ac.kr; hysong@bubble.yonsei.ac.kr; jhkim@bubble.yonsei.ac.kr.*

Manuscript received 30 May 1996; revised 2 Sept. 1997. Recommended for acceptance by T. Ishida.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 105668.

Therefore, entropy increases a lot more when the number of classes increases from two to three, for example, than when it increases from eight to nine, provided that the probability distribution of classes is uniform. One property of entropy of the partition is the following. If a partition induced on a set S by an attribute A' is a refinement of a partition by an attribute A, then the entropy $E(A', S)$ of the partition induced by A' is never higher than the entropy $E(A, S)$ of the partition induced by A, that is,

$$E(A', S) \le E(A, S),$$

where the equality holds if and only if the class distribution before and after the partition is maintained identically [8]. The point we would like to make here is that, as we refine further and further, the entropy of the partition decreases unless the class distribution before and after each partition is maintained identically, regardless of the appropriateness of distinguishing classes.

2.2 Two Criteria Based on Entropy

Quinlan defined gain as

$$\text{gain}(A,S) \triangleq \text{Ent}(S) - \text{E}(A,S) \tag{3}$$

and used it as a criterion in attribute selection for ID3 [1]. Because the entropy of the partition decreases as the partition is refined as mentioned in the previous subsection, the gain in (3) prefers the finer partition. In general, however, the decision tree generated only by the gain is known to be not only complex but also resulting in a poor performance in the classification rate.

In the enhanced version of ID3 called ID3-IV, a gain ratio is used, which is defined as follows [1]. Let an attribute A have values $a_1, a_2, ..., a_v$, and let the number of samples with value a_i of the attribute A be u_i for each $i = 1, 2, ..., v$, resulting in a probability distribution

$$q_i = \frac{u_i}{\sum_{j=1}^{v} u_j}$$

for each $i = 1, 2, ..., v$. A function IV on these probabilities is defined as

$$\text{IV}(A,S) \triangleq -\sum_{i=1}^{v} q_i \log_2 q_i \tag{4}$$

and gain ratio is defined as

$$\text{gain ratio}(A,S) \triangleq \frac{\text{gain}(A,S)}{\text{IV}(A,S)} \tag{5}$$

One advantage of using the gain ratio in (5) is the following. The value of IV increases logarithmically as the number of branches increases provided that the number of samples belonging to each branch is a constant. Therefore, we could expect that gain ratio less prefers a finer partition, because it is the gain divided by IV. Actually, gain ratio was shown to generate decision trees with improved performance over gain [9] and has been used as the attribute selection criterion in such well-known algorithms as GID3, GID3*, C4, C4.5, and so on. On the other hand, gain ratio increases as the value of IV decreases, provided that the gain remains the same. The value of IV decreases as the number of samples in each branch becomes less uniform. Therefore, as Mingers has observed [10], gain ratio has a tendency to prefer less uniform partition and, hence, generates a decision tree, regardless of the degree of classification, according only to the distribution of samples allocated to each branch. We illustrate all these criteria and their relation in Fig. 1 and the following two examples.

EXAMPLE 1. Let S consist of 10 Os and 10 Xs that are partitioned by an attribute A into three branches.

The values of gain and gain ratio are the following:

$$Ent(S) = -\frac{10}{20}\log_2\frac{10}{20} - \frac{10}{20}\log_2\frac{10}{20} = 1$$

$$E(A,S) = \frac{8}{20}\cdot 0 + \frac{5}{20}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{7}{20}\cdot 0 = 0.243$$

$$gain(A,S) = 1 - 0.243 = 0.757$$

$$IV(A,S) = -\frac{8}{20}\log_2\frac{8}{20} - \frac{5}{20}\log_2\frac{5}{20} - \frac{7}{20}\log_2\frac{7}{20} = 1.559$$

$$gain\ ratio(A,S) = \frac{0.757}{1.559} = 0.486$$

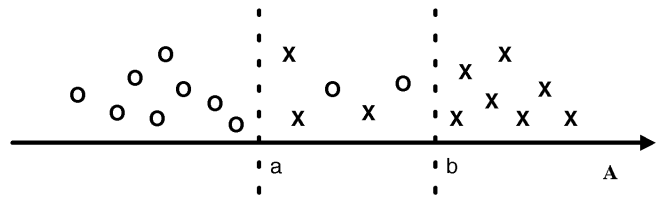

Fig. 1. A partition of a set of 20 samples into three subsets.

EXAMPLE 2. Let a refinement A' of A be applied to S in Example 1, resulting in a finer partition as shown in Fig. 2.

The values of gain and gain ratio are the following:

$$Ent(S) = 1$$

$$E(A',S) = \frac{8}{20}\cdot 0 + \frac{3}{20}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{2}{20}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{7}{20}\cdot 0 = 0.238 < E(A,S)$$

$$gain(A',S) = 1 - 0.238 = 0.762 > gain(A,S)$$

$$IV(A',S) = -\frac{8}{20}\log_2\frac{8}{20} - \frac{3}{20}\log_2\frac{3}{20} - \frac{2}{20}\log_2\frac{2}{20} - \frac{7}{20}\log_2\frac{7}{20} = 1.802$$

$$gain\ ratio(A',S) = \frac{0.762}{1.802} = 0.423 < gain\ ratio(A,S)$$

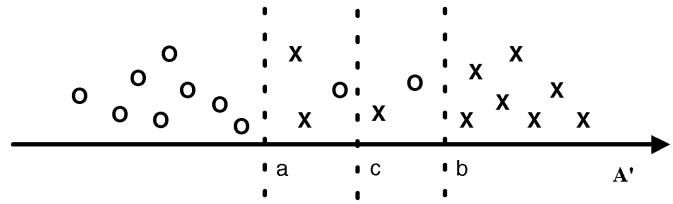

Fig. 2. A refined partition of a set of 20 samples into four subsets.

2.3 A New Criterion

The starting point is to reconsider the physical meaning of entropy used in both gain and gain ratio. As far as all of the authors are aware, the log base two has been used without any theoretical consideration in calculating the entropy, resulting in the criteria which are meaningful only in the "binary" partition, in some sense. However, in general, an attribute may cause an n-ary partition of the samples, and we argue that the log base n, for $n \ge 2$, should be used in order to better accommodate the characteristics of the attribute in the classification and discretization. Thus, we propose the new criterion below.

DEFINITION. *Let a sample set S be composed of k classes $c_1, c_2, ..., c_k$*

with probabilities $p_1, p_2, \ldots, p_k$, respectively. Let an attribute A partition S into n disjoint subsets $S_1, S_2, \ldots, S_n$. Denote

$$P_i \triangleq \frac{|S_i|}{|S|}, \quad i = 1, 2, \ldots, n,$$

and

$$P_i \triangleq \frac{|S_i \cap c_j|}{|S_i|}, \quad i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, k.$$

Then, define

$$\text{normalized gain}(A, S, n) \triangleq \left(-\sum_{j=1}^{k} p_j \log_n p_j\right) - \left(-\sum_{i=1}^{n} P_i \sum_{j=1}^{k} p_{ij} \log_n p_{ij}\right) \tag{6}$$

The normalized gain in the above definition is a "normalized" gain in the sense that the gain with log base two is divided by $\log_2 n$ and has an obvious relation

$$\text{normalized gain}(A, S, n) = \frac{\text{gain}(A, S)}{\log_2 n}, \quad n \geq 2.$$

In the remainder of this section, we will argue that, in some situations, the normalized gain is a better criterion than the gain or the gain ratio in the selection of an attribute in the process of generating a better decision tree. Roughly speaking, the use of the normalized gain can be interpreted as asking only the minimum expected number of *n*-ary questions in order to determine the outcome of an experiment governed by a random variable with probabilities $P_1, P_2, \ldots, P_n$. It is obvious that this number of *n*-ary questions becomes much smaller than the minimum expected number of binary questions to be asked for the same purpose. Specifically, we consider the following two situations summarized as theorems.

THEOREM 1. *Let a sample set S be made up of k classes $c_1, c_2, \ldots, c_k$ with probabilities $p_1, p_2, \ldots, p_k$, respectively. Let an attribute A (and B, respectively) partition S into m (and n, respectively) disjoint subsets. If each subset by A (and also by B) contains samples belonging to exactly one class, and if $n > m \geq 2$, then*

1) *gain(A, S) = gain(B, S);*
2) *normalized gain(A, S, m) > normalized gain(B, S, n); and*
3) *gain ratio(A, S) > gain ratio(B, S) if and only if IV(A, S) < IV(B, S).*

PROOF. Since each subset contains samples belonging to exactly one class,

$$E(A, S) = E(B, S) = 0.$$

Therefore,

$$\text{gain}(A, S) = \text{Ent}(S) = \text{gain}(B, S),$$

and

$$\begin{aligned} \text{normalized gain}(A, S, m) &= \frac{\text{Ent}(S)}{\log_2 m} \\ &> \frac{\text{Ent}(S)}{\log_2 n} \\ &= \text{normalized gain}(B, S, n) \end{aligned}$$

Finally, since $m \geq 2$, $\text{IV}(B, S) > 0$, and $\text{IV}(A, S) > 0$. Therefore,

$$\text{gain ratio}(A, S) = \frac{\text{Ent}(S)}{\text{IV}(A, S)} > \frac{\text{Ent}(S)}{\text{IV}(B, S)} = \text{gain ratio}(B, S)$$

if and only if $\text{IV}(A, S) < \text{IV}(B, S)$. □

EXAMPLE 3. A situation in which normalized gain works better than gain and gain ratio is shown in Fig. 3.

gain(A, S) = 0.918 = *gain*(B, S)
normalized gain(A, S, 2) = 0.459 > *normalized gain*(B, S, 3) = 0.395
IV(A, S) = 1.918 > *IV*(B, S) = 1.631
gain ratio(A, S) = 0.479 < *gain ratio*(B, S) = 0.563

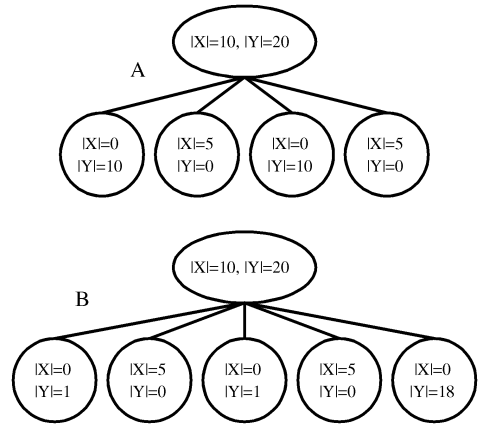

Fig. 3. Normalized gain is better than gain and gain ratio.

We would like to argue that the normalized gain is a better criterion than both gain and gain ratio in some situations satisfied by the assumptions in Theorem 1. First, normalized gain is always better than gain. Second, normalized gain is a better criterion than gain ratio if IV(*A*, *S*) > IV(*B*, *S*). We would like to argue that the condition IV(*A*, *S*) > IV(*B*, *S*) can be sometimes satisfied under the assumption of Theorem 1 even though $m < n$, as shown in Example 3. In the following theorem, we could think of another situation in which normalized gain and gain work better than gain ratio.

THEOREM 2. *Let a sample set S be made up of k classes $c_1, c_2, \ldots, c_k$ with probabilities $p_1, p_2, \ldots, p_k$, respectively. Let an attribute A and another attribute B both partition S into the same number, say $n \geq 2$, of disjoint subsets such that each subset by A contains samples belonging to exactly one class, and that at least one subset by B contains samples belonging to two or more classes. Then,*

1) *gain(A, S) > gain(B, S),*
2) *normalized gain(A, S, n) > normalized gain(B, S, n), and*
3) *gain ratio(A, S) < gain ratio(B, S) if and only if*

$$\frac{E(B, S)}{Ent(S)} < 1 - \frac{IV(B, S)}{IV(A, S)}. \tag{7}$$

PROOF. The first and the second assertions are obvious. For the third, recall that

$$\text{gain ratio}(A, S) = \frac{\text{gain}(A, S)}{\text{IV}(A, S)} = \frac{\text{Ent}(S) - \text{E}(A, S)}{\text{IV}(A, S)}.$$

Therefore,

$$\begin{aligned} &\frac{\text{Ent}(S) - \text{E}(A, S)}{\text{IV}(A, S)} < \frac{\text{Ent}(S) - \text{E}(B, S)}{\text{IV}(B, S)} \\ &\leftrightarrow \frac{\text{IV}(B, S)}{\text{IV}(A, S)} < \frac{\text{Ent}(S) - \text{E}(B, S)}{\text{Ent}(S) - \text{E}(A, S)} \\ &\leftrightarrow \frac{\text{IV}(B, S)}{\text{IV}(A, S)} < 1 - \frac{\text{E}(B, S)}{\text{Ent}(S)} \\ &\leftrightarrow \frac{\text{E}(B, S)}{\text{Ent}(S)} < 1 - \frac{\text{IV}(B, S)}{\text{IV}(A, S)} \end{aligned}$$

□

Some remarks should be followed on the conclusions of Theorem 2. First, in the situation satisfied by the assumptions in Theorem 2, normalized gain is again a better criterion than gain ratio, and both gain and normalized gain are equally good criteria. Second, one may be interested in the situation as to when the condition in (7) can be satisfied. It is easy to check that

$$0 < \frac{\mathrm{E}(B,S)}{\mathrm{Ent}(S)} < 1$$

because E(B, S) is the weighted average of the entropies of subsets of S, as defined in (2). Now, if IV(A, S) < IV(B, S), then the right-hand side of (7) becomes negative and the inequality (7) will not be satisfied at all. It means that the partition by the attribute B looks more like the uniform partition than that by A. On the other hand, if IV(A, S) > IV(B, S), that is, if the partition by the attribute A looks more like the uniform partition than that by B, the right-hand side of (7) becomes positive. Now, in order to satisfy the inequality (7) in this situation, the attribute B should give a partition which looks far from the uniform. One such case is illustrated in the example below.

EXAMPLE 4. A situation in which normalized gain works better than gain ratio is shown in Fig. 4. Here, normalized gain and gain both works the same.

$$gain(A, S) = 0.547 > gain\ (B, S) = 0.482$$
$$normalized\ gain\ (A, S, 3) = 0.345 > normalized\ gain\ (B, S, 3) = 0.304$$
$$IV(A, S) = 1.585 > IV(B, S) = 1.199$$
$$gain\ ratio(A, S) = 0.345 < gain\ ratio(B, S) = 0.402$$

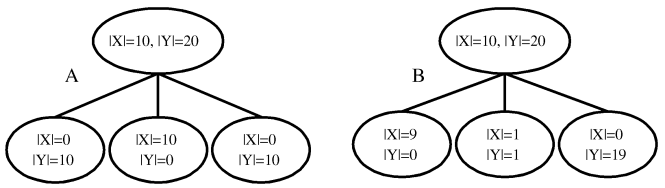

Fig. 4. Normalized gain is better than gain ratio.

3 SIMULATION

In order to show the efficiency of the proposed criterion, we should compare the performance of various classifiers using decision trees generated by the same algorithm except for the criterion. The process of generating recursively a decision tree for each dataset includes the following steps, in general:

- **(Step 1)** applying an attribute to all samples in the set and list up the attribute values;
- **(Step 2)** discretizing (or grouping) the samples according to the attribute values (→ This gives a partition on S); and
- **(Step 3)** calculating the value of a predetermined criterion and selecting the attribute which gives the maximum.

A well-known decision-tree generator C4.5 uses the same criterion in both discretization (second step) and selection (third step) of an attribute. However, C4.5 uses one of multiple different methods in the discretization (or grouping) according to the type of samples. The choice of a method in this step in C4.5 is somewhat optimized for the gain ratio, because it uses only the binary partition except for the type of samples with nonordered discrete attribute values. In order to guarantee a fair comparison, we believe that an n-ary partition should be allowed, and, hence, in our simulation, a predefined single algorithm (for discretizing and grouping) was applied throughout every dataset without pruning. We would like to leave as a future study the optimizing this algorithm including pruning relative to the proposed criterion. We would like to note that, in our simulation, the second step in the above process uses the same criterion as in the third step. The second step can be described in detail as follows:

- **(Step 2.1)** Take the initial partition D_N of samples by splitting at every class boundary. Let this give N intervals.
- **(Step 2.2)** For $n = N - 1, N - 2, N - 3, ..., 2$, repeat the following recursively: In the partition D_{n+1}, there are n ways to merge two adjacent intervals. Select the best partition scheme, call it D_n, relative to the given criterion among these n possibilities each of having n intervals.
- **(Step 2.3)** Select the best partition among $N - 1$ partitions $D_N, D_{N-1}, \ldots, D_2$ determined in Step 2.2.

Twelve different datasets obtained from the UCI Repository without any modification have been used for the experiment, Table 1 shows a summary of characteristics of the datasets, in which the upper seven sets consist of samples mainly with discrete attributes, and the remaining five sets consist of samples mainly with continuous attributes.

TABLE 1
DESCRIPTION OF DATASETS

Name of dataset	Characteristics			
	Number of classes	Number of attributes	Number of samples	attributes
Tic-Tac-Toe Endgame	2	9	958	mainly discrete
Fitting Contact Lenses	3	4	24	
Car	4	6	1728	
Nursery	5	8	12960	
Small Soybean	5	35	47	
Zoo	7	16	101	
Flag to Religion	8	23	194	
Ionosphere	2	34	351	mainly continuous
Iris Plants	3	4	150	
Wine Recognition	3	13	178	
Glass Identification	6	9	214	
Image Segmentation	7	19	2310	

In this experiment, a cross-validation [11] was performed as follows. The available data were divided into 10 blocks so as to make the number of samples and class distribution in each block as uniform as possible. Ten different classifiers were then built, each of which was based only on nine blocks and the resulting classifier was tested on the samples in the remaining block. For each of 10 classifiers, the three parameters of error rate, number of leaves, and weighted depth are obtained and then averaged. Here, the error rate measures the correctness of classification for the test data, the number of leaves measures the complexity of the resulting decision trees and is closely related to the classification rate [12]. The weighted depth is an average length of paths, each of which is weighted by the probability of samples allocated to each leaf, and it measures how fast the classification is performed.

Table 2 represents the average performance of classifiers which are generated by using gain, gain ratio, normalized gain, respectively, as a single criterion for both discretization and selection of attributes. As Table 2 clearly shows, we can see that normalized gain works better than gain or gain ratio for most of the datasets.

4 CONCLUSION

In this paper, we have proposed a new criterion, called normalized gain, for the selection and discretization of attributes to be used in the process of generating decision trees. The proposed criterion is not totally new in the sense that it is based on entropy, but it is new since it is a "normalized" version of the previously well-known criterion, the gain.

We have clearly demonstrated by two theorems those situations in which normalized gain works better than gain or gain ratio, and we have done some simulation to demonstrate that normalized gain works also better for a wide variety of datasets used in the simulation than gain and gain ratio. Since gain ratio works better in some situations as the simulation result shows, we can imagine that it happened in those situations not covered by the assumptions given in two theorems in Section 2, which seems to need a further study. We would like to emphasize that normalized gain works better than or at least equally as either gain or gain ratio does in any situation covered by the assumptions given in two theorems.

TABLE 2
RESULT OF SIMULATION

Dataset	Estimate	Criterion		
		gain	gain ratio	normalized gain
Tic-Tac-Toe	error rate (%)	31.4	30.4	19.1
	# of leaves	158.4	138.0	77.2
	weighted depth	4.5	4.5	5.0
Lenses	error rate (%)	26.7	23.3	23.3
	# of leaves	6.8	6.7	6.7
	weighted depth	2.2	2.2	2.2
Car	error rate (%)	10.8	11.6	2.1
	# of leaves	160.4	160.7	77.8
	weighted depth	2.9	3.2	4.0
Nursery	error rate (%)	1.3	1.2	0.1
	# of leaves	412.6	433.2	197.9
	weighted depth	3.4	3.8	5.0
Soybean	error rate (%)	1.7	1.7	0.0
	# of leaves	5.0	4.0	4.0
	weighted depth	1.3	1.8	2.0
Zoo	error rate (%)	4.5	1.8	3.6
	# of leaves	13.8	10.5	9.9
	weighted depth	2.3	3.0	2.7
Flag	error rate (%)	71.0	67.8	69.6
	# of leaves	120.7	101.8	94.8
	weighted depth	3.9	8.4	6.4
Ionosphere	error rate (%)	38.5	23.0	19.9
	# of leaves	81.6	41.6	29.8
	weighted depth	1.2	2.2	2.9
Iris	error rate (%)	16.3	11.9	10.0
	# of leaves	15.4	11.7	8.4
	weighted depth	1.2	1.7	2.6
Wine	error rate (%)	41.5	46.2	15.4
	# of leaves	51.7	51.2	7.9
	weighted depth	1.1	1.2	2.8
Glass	error rate (%)	76.0	65.3	53.6
	# of leaves	117.9	104.0	59.1
	weighted depth	1.1	1.9	3.1
Image	error rate (%)	55.6	30.5	9.0
	# of leaves	816.1	422.8	72.6
	weighted depth	1.1	3.2	4.3

Finally, optimization of discretizing or grouping algorithm specifically for the proposed criterion should be studied in the near future, so that a comparison can be done between the decision-tree generator with optimum discretizing or grouping algorithm for the proposed criterion and that with (possibly different) optimum algorithm for other well-known criterion.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful suggestions in improving the earlier draft of this paper. This research was supported by Korea Telecom Research and Development Group under Contract 96-22.

REFERENCES

[1] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.
[2] J. Cheng, U.M. Fayyad, K.B. Irani, and Z. Quian, "Improved Decision Trees: A Generalized Version of ID3," *Proc. Fifth Int'l Conf. Machine Learning*. San Mateo, Calif.: Morgan Kaufmann, pp. 100-108, 1988.
[3] U.M. Fayyad, "Branching on Attribute Values in Decision Tree Generation," *AAAI-94; Proc. 12th Nat'l Conf. Artificial Intelligence*, pp. 601-606, Seattle, Wash., 31 July–4 Aug. 1994.
[4] J.R. Quinlan, P.J. Compton, K.A. Horn, and L. Lazarus, "Inductive Knowledge Acquisition: A Case Study," J.R. Quinlan, *Applications of Expert Systems*. Reading, Mass.: Addison-Wesley, pp. 157-173, 1987.
[5] J.R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, Calif.: Morgan Kaufmann, 1993.
[6] R.B. Ash, *Information Theory*. New York: Interscience, 1965.
[7] J.S. Lim, *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, N.J.: Prentice Hall, 1990.
[8] U.M. Fayyad, *On the Induction of Decision Trees for Multiple Concept Learning*, PhD dissertation, Univ. of Michigan, 1991.
[9] J.R. Quinlan, "Decision Trees and Multi-Valued Attributes," J. Richard, ed., *Machine Intelligence*, vol. 11. Oxford, England: Oxford Univ. Press, pp. 305-318, 1988.
[10] J. Mingers, "An Empirical Comparison of Selection Measures for Decision-Tree Induction," *Machine Learning*, vol. 3, no. 4, pp. 319-342, 1989.
[11] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*. Belmont, Calif.: Wadsworth, 1984.
[12] U.M. Fayyad and K.B. Irani, "What Should Be Minimized in a Decision Tree?" *AAAI-90; Proc. Eighth Nat'l Conf. Artificial Intelligence*, pp. 749-754, 1990.