RC 20072 (88792) 5/23/95
Computer Science/Mathematics 13 pages

Research Report

Use of Randomization to Normalize Feature Merits

Se June Hong, J. R. M. Hosking and Shmuel Winograd

IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598

IBM **Research Division**
Almaden • T.J. Watson • Tokyo • Zurich

Use of randomization to normalize feature merits

Se June Hong, J. R. M. Hosking and Shmuel Winograd

Mathematical Sciences Dept.
IBM Research Division
P. O. Box 218
Yorktown Heights, NY 10598

Abstract. Feature merits are used for feature selection in classification and regression as well as for decision tree generation. Commonly used merit functions exhibit a bias towards features that take a large variety of values. We present a scheme based on randomization for neutralizing this bias by normalizing the merits. The merit of a feature is normalized by division by the expected merit of a feature that is random noise taking the same distribution of values as the given feature. The noise feature is obtained by randomly permuting the values of the given feature. The scheme can be used for any merit function including the Gini and entropy measures. We demonstrate its effectiveness by applying it to the contextual merit defined by Hong (*IBM Res. Rep. RC19664*, 1994).

Topic Area: Knowledge Mining and Discovery, Data Abstraction and Modeling.

Contact: Se June Hong
HONG@watson.ibm.com
(914) 945-2265 (3434 fax)

1. Introduction

In developing decision models for classification or regression, one of the primary concerns is which of the available features or variables should be used in building the model. For this purpose, one usually computes one of several flavors of merit (or demerit) values for each feature. In generating decision tree models, one recursively splits the current decision node by the test feature that has the highest merit for the subset of training examples at the node. Most popular merits are based on the Gini function as in CART (Breiman et al., 1984), or on the entropy function as in C4.5 (Quinlan, 1993). Both approaches suffer from an inherent bias that favors test features that split into more branches. This is also true of the new contextual merit function proposed in Hong (1994). White and Liu (1994) have convincingly demonstrated the prevalence of this bias and discussed undesirable effects that it has on the models generated from such biased merits.

When a feature variable takes many distinct values it certainly has more power to model the class or the response variable, whether the values are symbolic (categorical) or numeric. At one extreme, a feature whose value is distinct for all the training examples is by itself sufficient to model the class or the response variable. This can happen quite often in real data, in the form of names, account numbers, or even the example ID number. The model generated using such a feature enjoys perfect precision, but its recall rate for the future examples may well be zero. Hong used the term "variety effect" to describe this phenomenon especially in connection with numerical features that have mostly unique values for each example.

This observation has to do with our understanding of the domain, exogenous to the example data at hand. In an attempt to address this problem Quinlan divides the information measure between the feature and class by the entropy of the feature itself, which tends to be larger when there is more variety.

We claim that there is a domain independent rationale in addressing this problem as well as some other concerns. We ask what the merit (however computed) would be for a random feature that had the same distribution of values as the given feature. If the merit of the original feature is close to or less than the average merit of random features with the same distribution, the feature should not be assigned a high merit value. One can at will create arbitrarily distributed random features and insert them in the data, and a feature that does no better on the average than a chance feature might as well be replaced with a random one. Hence, we propose to compute the average merit of random features with the same distribution as the given one, and use it to normalize the original merit. It will be shown that, in most cases, this normalizing factor can be easily computed. At worst, we can obtain a simulated average value by trying several randomized features. For reasonably large data sets, we have found that such simulated average values are very close to the full average, most of the time.

A simple way to obtain a random feature with the same distribution as the given feature is to randomly permute the values of the given feature itself among the examples. An observed quantity, such as the impurity of the distribution of the class values given the feature values, may then be compared with the distribution of values that the quantity would take under random permutation of the feature values. This is the basis of randomization tests, widely used in statistics (Edgington, 1986). The randomization "itself represents a hypothesis of interest which permits the statistical evaluation of apparent associations for the experience of a fixed population without any underlying probability distribution assumptions" (Koch, 1982).

Once normalized, Gini- or entropy-based merits no longer favor high branching features. In particular, a feature that takes a unique value for every example would have the same merit measure as its random counterpart and would therefore have normalized merit of one. A further benefit of normalization for contextual merits is that by allowing for the "variety effect" it obviates the need to calibrate the merits of numeric features against those of symbolic features. We present experimental results on this in section 4.

2. Normalizing Gini- and entropy-based merits

We denote by $I(p_1, \ldots, p_m)$ the impurity of a set of probabilities $p_1, p_2, \ldots, p_m$, $p_1 + p_2 + \ldots + p_m = 1$. An impurity measure should satisfy $I(p_1, \ldots, p_m) = 0$ whenever $p_i = 1$ for some i, and should be maximized when $p_i = 1/m$ for all i. Reasonable forms for $I(p_1, \ldots, p_m)$ include the Gini measure

$$G(p_1, \ldots, p_m) = 1 - \sum_i p_i^2 \tag{1}$$

and the entropy measure

$$H(p_1, \ldots, p_m) = -\sum_i p_i \log p_i \,. \tag{2}$$

In a categorical classification problem, there are N examples over which the dependent variable ("class") takes C distinct values with frequencies b_i, $i = 1, \ldots, C$. We consider an explanatory variable ("feature") that takes F distinct values with frequencies a_j, $j = 1, \ldots, F$. Although we consider a feature that branches into an arbitrary number of values, F, common practice has been to use either the full branching corresponding to all unique values of the feature, or separate equality testing (binary) sub-features, in the case of symbolic features. A more general branching can be induced from some useful partition of the feature values. When the feature is numeric, most decision tree generation methods use only the binary branching of Greater-Than test-features derived from the unique values of the feature.

The conjunction of class i and feature value j occurs with frequency x_{ij}. The occurrences of class and feature values can be summarized by a contingency table, as in Table 1.

Table 1. Contingency table for feature $\times$ class cross-classification.

		Feature value				
		1	2	...	F	Total
Class	1	x_{11}	x_{12}	...	x_{1F}	b_1
	2	x_{21}	x_{22}	...	x_{2F}	b_2
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	C	x_{C1}	x_{C2}	...	x_{CF}	b_C
	Total	a_1	a_2	...	a_F	N

The impurity of the class variable is denoted by

$$I(C) = I(b_1/N, \ldots, b_C/N)\,. \tag{3}$$

For those examples corresponding to a particular value of the feature, the impurity is denoted by

$$I(C|F=j) = I(x_{1j}/a_j, \ldots, x_{Cj}/a_j)\,. \tag{4}$$

When feature value j occurs with probability p_j the average impurity of the class variable, given the feature, is

$$I(C|F) = \sum_{j=1}^{F} p_j\, I(x_{1j}/a_j, \ldots, x_{Cj}/a_j) \tag{5}$$

For the set of N examples the proportion of occurrences of feature value j is a_j/n, and using this value as p_j in (5) yields

$$I(C|F) = \sum_{j=1}^{F} \frac{a_j}{N}\, I(x_{1j}/a_j, \ldots, x_{Cj}/a_j)\,. \tag{6}$$

This is the impurity that remains in the class variable after the information present in the feature variable has been used, i.e. branching due to the test outcome.

When using Gini's measure of impurity, (6) becomes

$$G(C|F) = \sum_{j=1}^{F} \frac{a_j}{N} \left\{ 1 - \sum_{i=1}^{C} \left(\frac{x_{ij}}{a_j} \right)^2 \right\} = 1 - \frac{1}{N} \sum_{j=1}^{F} \sum_{i=1}^{C} \frac{x_{ij}^2}{a_j}\,. \tag{7}$$

The best feature is the one that achieves the lowest value of (7).

When using the entropy measure of impurity, (6) becomes

$$H(C|F) = \sum_{j=1}^{F} \frac{a_j}{N} \sum_{i=1}^{C} -\frac{x_{ij}}{a_j} \log \frac{x_{ij}}{a_j} = \frac{1}{N} \left(\sum_{j=1}^{F} a_j \log a_j - \sum_{j=1}^{F} \sum_{i=1}^{C} x_{ij} \log x_{ij} \right). \tag{8}$$

The best feature is the one that achieves the lowest value of (8), or, equivalently, the highest value of the "mutual information" $H(C) - H(C|F)$.

We wish to compare different feature variables in terms of their ability to explain the variation in the class variable. However, a direct comparison of values of $I(C|F)$ for different features unduly favors features that take a large number of distinct values. In the extreme case, a feature that takes N distinct values for the N examples achieves complete discrimination between different classes, giving $I(C|F) = 0$, even though the feature may consist of random noise and be useless for predicting the classes of future examples. For a fair assessment of the observed value of $I(C|F)$ for a feature we seek to compare it with the value of $I(C|F)$ that would be achieved, on average, by a random-noise feature taking the same number of distinct values, with the same frequencies, as the given feature. We now compute what this average value would be.

A random noise feature may be obtained by randomly permuting the values of any given feature variable. Consider a feature variable $\mathcal{F}$ that, for the N total examples, takes F distinct values with frequencies a_j, $j = 1, \ldots, F$. A random permutation of the N values of $\mathcal{F}$ yields a feature $\mathcal{F}^*$ that takes the same values as $\mathcal{F}$ with the same frequencies, but whose values are completely uncorrelated with those of the class variable $\mathcal{C}$:

$$\Pr[\mathcal{F}^* = j \,|\, \mathcal{C} = i] = a_j/N \tag{9}$$

regardless of the value of i.

The values taken by the random-noise feature may also be represented by the cross-classification in Table 1, but with the cell counts x_{ij} being random variables, which we denote by X_{ij}. Under random permutation, any individual X_{ij} has a hypergeometric distribution, with

$$\Pr[X_{ij} = k] = \binom{b_i}{k}\binom{N - b_i}{a_j - k} \bigg/ \binom{N}{a_j}. \tag{10}$$

From this distribution we can calculate the expected values needed to computed the expected values of $G(C|F)$ and $H(C|F)$.

For the Gini impurity measure, we use standard results for the hypergeometric distribution:

$$EX_{ij} = \frac{a_j b_i}{N}, \qquad \operatorname{var} X_{ij} = \frac{a_j(N - a_j)b_i(N - b_i)}{N^2(N-1)} \tag{11}$$

(Stuart and Ord, 1987, p. 167), whence

$$EX_{ij}^2 = (EX_{ij})^2 + \operatorname{var} X_{ij} = \frac{a_j b_i}{N(N-1)}(N - a_j - b_i + a_j b_i). \tag{12}$$

Thus

$$E\{G(C|F)\} = 1 - \frac{1}{N}\sum_{j=1}^{F}\sum_{i=1}^{C}\frac{b_i}{N(N-1)}(N - a_j - b_i + a_j b_i), \tag{13}$$

which, using the summations $\sum_j 1 = F$, $\sum_i 1 = C$ and $\sum_j a_j = \sum_i b_i = N$, reduces to

$$E\{G(C|F)\} = \frac{N-F}{N^2(N-1)}\Big(N^2 - \sum_{i=1}^{C} b_i^2\Big) \tag{14}$$

$$= \Big(\frac{N-F}{N-1}\Big)\, G(C)\,. \tag{15}$$

For the entropy-based impurity measure we must calculate $E(X_{ij}\log X_{ij})$. The exact value of the expectation is complicated, but a simple approximation can be derived for the case in which N is large and none of the a_j or b_i is small. The approximation is valid asymptotically as $N \to \infty$ with $a_j/N \to \alpha_j > 0$, $j = 1, \ldots, F$, and $b_i/N \to \beta_i > 0$, $i = 1, \ldots, C$. In the stochastic Taylor-series expansion of a function f of a random variable X about its mean μ,

$$f(X) \sim f(\mu) + f'(\mu)(X-\mu) + \tfrac{1}{2}f''(\mu)(X-\mu)^2 + \ldots\,, \tag{16}$$

we ignore terms after the third and take expectations, giving

$$E\{f(X)\} \approx f(\mu) + \tfrac{1}{2}f''(\mu)\,\mathrm{var}\,X\,. \tag{17}$$

We take $f(x) = x\log x$ and X to be X_{ij}/a_j. From (11), using the given asymptotics as $N \to \infty$, we have

$$E(X_{ij}/a_j) \sim \beta_i\,, \qquad \mathrm{var}(X_{ij}/a_j) \sim \frac{(1-\alpha_j)\beta_i(1-\beta_i)}{N\alpha_j}\,; \tag{18}$$

substituting into (17) we obtain the approximation

$$E\Big(\frac{X_{ij}}{a_j}\log\frac{X_{ij}}{a_j}\Big) \approx \beta_i\log\beta_i + \frac{(1-\alpha_j)(1-\beta_i)}{2N\alpha_j}\,, \tag{19}$$

whence

$$E\{H(C|F)\} \approx -\sum_{j=1}^{F}\alpha_j\sum_{i=1}^{C}E\Big(\frac{X_{ij}}{a_j}\log\frac{X_{ij}}{a_j}\Big) \tag{20}$$

$$\approx -\sum_{j=1}^{F}\beta_i\log\beta_i - \frac{1}{2N}\sum_{j=1}^{F}(1-\alpha_j)\sum_{i=1}^{C}(1-\beta_i) \tag{21}$$

$$\approx H(C) - \frac{1}{2N}(F-1)(C-1)\,. \tag{22}$$

When some of the frequencies a_j are small, a more accurate approximation to $E\{H(C|F)\}$ can be obtained from (20) by evaluating $E\{(X_{ij}/a_j)\log(X_{ij}/a_j)\}$ exactly for feature values for which a_j is small and using the approximation (19) when a_j is larger than some threshold A. Denoting by $\#\{j : a_j = a\}$ the number of distinct feature values that occur exactly a times in the N examples, and by $n^{[k]}$ the factorial power

$$n^{[k]} = n!/(n-k)! = n(n-1)\dots(n-k+1)\,, \tag{23}$$

we obtain

$$\begin{aligned} E\{H(C|F)\} \approx H(C) \sum_{j:a_j>A} \alpha_j + \frac{1}{2N}(C-1) \sum_{j:a_j>A} (1-\alpha_j) \\ + \frac{1}{N} \sum_{a=2}^{A} \#\{j : a_j = a\} \left\{ \sum_{k=2}^{a-1} \binom{a}{k} \sum_{i=1}^{C} \frac{b_i^{[k]}(N-b_i)^{[k]}}{N^{[k]}} (-k \log k) \right. \\ \left. + \left(1 - \sum_{i=1}^{C} \frac{b_i^{[a]}}{N^{[a]}}\right) a \log a \right\}. \end{aligned} \tag{24}$$

It is noteworthy that both (15) and (22) involve only F, the number of distinct values taken by the feature, and not the relative frequencies of the different values. Thus, for example, when only comparisons between binary features are of interest, comparing the observed merit with (15) or (22) will not affect the relative merits of the features. This is the case when using CART with branching on whether a symbolic feature is equal to a particular value or belongs to a particular set of values, or on whether a numeric feature is greater than a particular threshold.

There remains the question of how the expected value $E\{I(C|F)\}$ should be used to normalize an observed value $I(C|F)$. We suggest that the normalized impurity be defined by the ratio $I(C|F)/E\{I(C|F)\}$, or by 1 when $I(C|F) = E\{I(C|F)\} = 0$. This immediately identifies features that are no better than random, and gives good results when used with contextual merits, as shown by the example in section 4.

Other normalizations are possible. White and Liu (1994) evaluated several merit measures and found that those least subject to bias induced by the variety effect were, in their terminology, "probability-based", i.e. constructed as the significance level of a statistical hypothesis test of whether the observed value of a particular impurity measure was consistent with the feature's being random noise. We could construct such a measure in the present context, using the distribution of $I(C|F)$ under random permutations of the values of the feature variable as the null distribution with respect to which the significance level of an observed value of $I(C|F)$ would be computed.[1] We feel, however, that it is important to retain as much as possible of the initial idea of comparing $I(C|F)$ values of different features, while allowing for the variety effect, and that this is most simply achieved by using the ratio $I(C|F)/E\{I(C|F)\}$.

Quinlan (1993) advocated the normalized merit $\{H(C) - H(C|F)\}/H(F)$. This normalization is somewhat arbitrary and we feel that to achieve distinction between useful features and those that are no better than random noise, a normalization based on the expected impurity (8) is more appropriate. This suggests the use of $\{H(C) -$

[1] White and Liu's simulation experiments used a different null distribution, which did not preserve the observed frequencies a_j of the different values of the feature variable.

$H(C|F)\}/[H(C) - E\{H(C|F)\}]$ as a merit or, more simply, $H(C|F)/E\{H(C|F)\}$ as a demerit.

3. Normalizing contextual merits

Contextual merit, introduced by Hong (1994), assigns merit to a feature taking into account the degree to which other features are capable of discriminating between the same examples as the given feature. As an extreme instance, if two examples in different classes differ in only one feature, then that feature is particularly valuable—if it were dropped from the set of features, there would be no way of distinguishing the examples—and is assigned additional merit.

To define contextual merit, we first define the distance $d_{rs}^{(k)}$ between the values z_{kr} and z_{ks} taken by feature k for examples r and s. If the feature is symbolic, taking only a discrete set of values, we define

$$d_{rs}^{(k)} = \begin{cases} 0 & \text{if } z_{kr} = z_{ks}, \\ 1 & \text{otherwise.} \end{cases} \tag{25}$$

If the feature is numeric, we set a threshold t_k and define

$$d_{rs}^{(k)} = \min(|z_{kr} - z_{ks}|/t_k, 1)\,. \tag{26}$$

The distance between examples r and s is now defined to be

$$D_{rs} = \sum_{k=1}^{N_f} d_{rs}^{(k)}\,, \tag{27}$$

N_f being the number of features. When a particular feature f is of interest, we also write

$$D_{rs} = d_{rs}^{(f)} + \Delta_{rs}^{(f)}\,, \tag{28}$$

where

$$\Delta_{rs}^{(f)} = D_{rs} - d_{rs}^{(f)} = \sum_{k \neq f} d_{rs}^{(k)}\,. \tag{29}$$

The merit of feature f is now defined as

$$M_f = \sum_{r=1}^{N} \sum_{s \in \bar{C}(r)} w_{rs}^{(f)} d_{rs}^{(f)}\,, \tag{30}$$

where N is the number of examples, $\bar{C}(r)$ is the set of examples not in the same class as example r, and $w_{rs}^{(f)}$ is a weight function chosen so that examples that are close together, i.e. that differ in only a few of their features, are given greater influence in determining each feature's merit.[2] Hong (1994) used weights $w_{rs}^{(f)} = 1/D_{rs}^2$ if s is

[2]Small modifications of the definition of M_f, such as the "CM" algorithm used by Hong (1994), appear to make little difference in practice.

one of the k nearest neighbors to r, in terms of D_{rs}, in the set $\bar{C}(r)$, and $w_{rs}^{(f)} = 0$ otherwise; the number of nearest neighbors used by Hong was the logarithm (to base 2) of the number of examples in the set $\bar{C}(r)$. We make a modification that is convenient, and in practice almost equivalent: we define nearest neighbors in terms of $\Delta_{rs}^{(f)}$ and set the weights for the nearest neighbors to be $w_{rs}^{(f)} = 1/(1+\Delta_{rs}^{(f)})^2$. This yields a simpler expression for the expected merit of a random feature, which we now derive.

As before, we define a random-noise feature analogous to feature f by randomly permuting the values taken by f, viz. z_{fr}, $r = 1, \ldots, N$. Consider a permutation of $\{1, \ldots, N\}$ that sends r' to r and s' to s. Its merit is

$$\sum_r \sum_s d_{r's'}^{(f)}/(1+\Delta_{rs}^{(f)})^2 \ : \tag{31}$$

observe that our chosen weight function is not changed when the values of feature f are permuted. The expected merit under random permutation is

$$E(M_f) = \frac{1}{N!} \sum_{r=1}^{N} \sum_{s\in\bar{C}(r)} \sum_{r'=1}^{N} \sum_{s'=1}^{N} N(r,s,r',s') \frac{d_{r's'}^{(f)}}{(1+\Delta_{rs}^{(f)})^2} \,, \tag{32}$$

where $N(r,s,r',s')$ is the number of permutations of $\{1, \ldots, N\}$ that send r' to r and s' to s, and is equal to $(N-2)!$ if $r' \neq s'$ and to 0 if $r' = s'$. We therefore have

$$\begin{aligned} E(M_f) &= \frac{1}{N(N-1)} \sum_{r=1}^{N} \sum_{s\in\bar{C}(r)} \sum_{\substack{r'=1 \\ r'\neq s'}}^{N} \sum_{s'=1}^{N} \frac{d_{r's'}^{(f)}}{(1+\Delta_{rs}^{(f)})^2} \\ &= \Bigg(\sum_{r=1}^{N} \sum_{s\in\bar{C}(r)} \frac{1}{(1+\Delta_{rs}^{(f)})^2} \Bigg) \Bigg(\frac{1}{N(N-1)} \sum_{\substack{r'=1 \\ r'\neq s'}}^{N} \sum_{s'=1}^{N} d_{r's'}^{(f)} \Bigg) \,. \end{aligned} \tag{33}$$

Expression (33) is reasonably convenient for computation. The second term in (33) is the average distance between the feature-f values of two randomly chosen examples, and can be further simplified for a symbolic feature. For a symbolic feature that takes F distinct values with frequencies a_j, $j = 1, \ldots, F$, this average distance is

$$\frac{1}{N(N-1)} \sum_j \sum_{k\neq j} a_j a_k = \frac{1}{N(N-1)} \sum_j a_j(N-a_j) \tag{34}$$

$$= \frac{N}{N-1} \Big(1 - \sum_j (a_j/N)^2 \Big) \tag{35}$$

$$= \frac{N}{N-1} G(F) \tag{36}$$

$$\approx G(F) \,, \tag{37}$$

where $G(F)$ is the Gini measure of impurity of the feature values.

4. Effect of normalized contextual merits

Hong (1994) used classification problems derived from Exclusive-Or functions to demonstrate the effectiveness of contextual merits.[3] Let us briefly describe these problems. The truth table for an Exclusive-Or function of n variables contains 2^n rows, corresponding to the binary n-tuple values, and $n+1$ columns, the last of which constitutes the class. The class value is 0 if the n-tuple contains an even number of 1s and 1 otherwise, which gives it another common name, the odd function. For given N, the number of examples, we replicate this truth table until the size just exceeds $2N$. We randomly select N rows from this replicated table, then split the n-tuple feature portion from the last column of class, and insert an $N \times m$ array of random binary values. The resulting table is defined as EXOR(n, m, N). It was shown in Hong (1994) that the merits of the first n features are higher than those of the m remaining random features, provided that there is a sufficient number of examples N to support the Exclusive-Or function of n variables in the presence of m random noise variables.

When one converts the feature values of an EXOR problem and replaces them with their numericized counterparts, such merit differentials quickly diminish. The numericizing is carried out by a procedure called RANNUM which converts the symbolic value 0 into a random integer in the range of 0 to $R-1$ and 1 into a random number in the range of R to $2R-1$, where R is an integer parameter. The problem becomes even more serious when only a subset of the features are numericized. In this case the features that are numericized tend to have higher contextual merit values than do the unchanged features, regardless of whether they were among the first n EXOR variables or the random features. This phenomenon is attributed largely to the "variety effect", but it includes some of the "ramp effect" due to the ramp function used for numeric feature distance. One of the measures to counteract this in Hong (1994) was to calibrate the merits of numerical features based on external knowledge about comparability of numeric and symbolic features. Normalizing the contextual merits by the expected merit of their random counterparts automatically accomplishes the calibration.

We demonstrate this on an EXOR$(3, 10, 200)$ function and its derived problem where some of the features are numericized. In Cases A, B, and C below, X1, X2, and X3 denote the three Exclusive-Or variables and R1-R10 the random variables. CM0 denotes the contextual merit values as defined by Hong (1994), and RM0 is the average of the original contextual merit values of 10 randomized simulations for the feature. NCM0 denotes the normalized values, i.e. CM0/RM0. CM1 and RM1 are the corresponding values computed by the modifications introduced in the previous section with $w_{rs}^{(f)} = 1/(1+\Delta_{rs})^2$, and using (33) for the expected randomized merits.[4] Again, NCM1 is the normalized value, CM1/RM1. When the merits of X1-X3 are

[3] We study EXOR and numericized versions of it because other methods fail on these problems, and the ability to handle them is readily found to extend to many other benchmark and real problems.

[4] We present contextual merits CM0 and CM1 to demonstrate that many similar varieties of

Table 2. Merits of feature variables in EXOR problems.

Case A: EXOR(3, 10, 200).

f	CM0	RM0	NCM0	CM1	RM1	NCM1
X1	23.17	14.67	1.58	33.45	24.36	1.37*
X2	23.30	14.82	1.57*	34.36	23.95	1.43
X3	23.04*	14.58	1.58	33.38*	23.91	1.40
R1	12.98	11.64	1.12	22.69	21.12	1.07
R2	14.03#	12.44	1.13#	24.23#	21.68	1.12#
R3	10.72	10.97	0.98	21.30	20.63	1.03
R4	11.74	11.40	1.03	21.75	21.09	1.03
R5	12.58	12.40	1.01	22.93	21.29	1.08
R6	11.98	11.65	1.03	21.34	20.96	1.02
R7	11.78	12.20	0.97	22.34	21.32	1.05
R8	9.87	11.40	0.87	20.50	20.64	0.99
R9	12.53	11.98	1.05	22.55	21.09	1.07
R10	9.94	10.77	0.92	19.66	20.56	0.96
Ranking successful?	yes		yes	yes		yes

Case B: EXOR(3, 10, 200), modified by converting X3, R1-R4 into numbers through RANNUM (R=100), numeric distance threshold equal to 1/2 the range.

f	CM0	RM0	NCM0	CM1	RM1	NCM1
X1	9.40*	7.37	1.27	17.48*	14.74	1.19*
X2	10.75	7.63	1.41	19.21	14.89	1.29
X3	22.11	18.54	1.19*	24.01	19.71	1.22
R1	16.60	16.37	1.01	18.20	17.62	1.03
R2	17.83*	17.24	1.03	19.57#	18.49	1.06
R3	15.76	16.11	0.98	18.69	18.50	1.01
R4	16.60	16.30	1.02	17.64	17.91	0.98
R5	7.38	6.76	1.09#	14.72	14.14	1.04
R6	7.37	6.79	1.09	15.32	14.24	1.08#
R7	6.26	6.18	1.01	14.57	14.07	1.04
R8	6.12	5.99	1.02	13.68	13.91	0.98
R9	5.60	5.70	0.98	14.01	13.91	1.01
R10	6.76	6.38	1.06	14.99	14.16	1.06
Ranking successful?	no		yes	no		yes

Table 2, continued.

Case C: Same as Case B, but with numeric distance threshold equal to 1/3 the range.

f	CM0	RM0	NCM0	CM1	RM1	NCM1
X1	7.77*	6.11	1.27	13.43*	11.72	1.15
X2	8.12	6.11	1.33	14.36	11.77	1.22
X3	19.01	17.09	1.11*	21.12	19.09	1.11*
R1	16.53	16.21	1.02	18.37	18.16	1.01
R2	16.97#	17.08	0.99	19.16#	18.72	1.02
R3	15.26	16.25	0.94	17.68	18.14	0.97
R4	15.85	16.33	0.97	17.98	18.26	0.98
R5	5.34	5.76	0.93	11.24	11.26	1.00
R6	5.53	5.44	1.02	10.84	11.27	0.96
R7	5.81	5.33	1.09#	11.79	11.41	1.03#
R8	4.65	5.49	0.85	10.34	11.16	0.93
R9	4.83	5.14	0.94	10.62	11.20	0.95
R10	4.97	5.17	0.96	10.87	11.24	0.97
Ranking successful?	no		yes	no		yes

higher than those of R1-R10, we say that the merit scheme succeeded, which is indicated in the Cases below. In each case, to facilitate inspection, the smallest merit value among X1-X3 is marked by *, and the largest among R1-R10 by #.

For the all-symbolic Case A, CM0 and CM1, as well as their normalized merits, NCM0 and NCM1, all succeed. In Case B some of the features are numericized. The numericized features exhibit higher merits in both the CM0 and CM1 schemes, as noted by Hong (1994). However, the normalized merits, NCM0 and NCM1, both succeed in removing the bias introduced by the now high variety numericized features. Case C demonstrates the robustness of method to the choice of numerical distance threshold. Our usual choice is 1/2 of the range of the numerical variable. As can be seen, taking the threshold to be 1/3 of the range does not affect the qualitative outcome.

The effects of normalization shown in Cases B and C are typical of such problems derived from EXOR functions. Most practical problems contain mixed numeric and symbolic features, and the normalization clearly places their contextual merits on an equal footing. We have reason to believe that the normalization negates not only the variety effect, but also some of the other undesirable consequences of the "ramp effect" in computing the contextual merit of mixed features.

the weight function $w_{rs}^{(f)}$ for computing the contextual merit all produce similar qualitative results, although the exact properties of different variations need further study.

5. Concluding remarks

We have proposed a new way of removing the bias in feature merits that arises from the variety effect. For a variety of merit (or demerit) schemes, we consider what happens when the feature is replaced with a random feature having the same distribution of values. The expected value of merits computed for such random analogues of a given feature certainly begs a comparison to the merit of the feature itself. A simple way to use this information is to normalize the original merit by dividing by the expected merit of the random feature. A feature whose merit is lower than would be expected for a random feature is a candidate to be removed. Further, the normalized merit, which measures how much better a feature performs than its random counterpart, can be used to rank the features in place of the original merits.

The expected values of the Gini- and entropy-based merits for random features are easily computed. Normalization by division serves to remove the "variety effect" in these impurity measures. Many other normalizations are possible: this is a subject for further study. While Gini-based merit or demerit schemes benefit from normalization, it was seen that only the branching factor (the variety) matters. Hence if only binary branches are of interest, as is usual for example with CART, the binary decision tree will not be affected at all. In entropy-based schemes the number of branches is the most important contributor to the expected merit of a random feature, while the effect of the distribution of the feature values is largely negligible except for branches that occur with very low frequency; this is pertinent for the ID3/C4.5 family of tree generators.

The same notion carries over to the contextual merit. The random counterpart of contextual feature merit, which is also easily computed, is effectively used to self-calibrate the differences in merits of numeric and symbolic features. We mention here that the contextual merits are relative quantities in that the relative value indicates the discriminating power of a feature strictly in the context of the other features. When the simply normalized value is less than one, it means only that in the presence of all other features, a random feature would contribute more for the classification than the given feature. Thus in ranking the features one needs to balance "how much better than random" with "how much power to begin with", especially when the purpose is to remove less relevant features from a large pool of features. If a feature has very low contextual merit to begin with, whether it is much superior to its random counterpart is immaterial. There may therefore be better ways of normalizing than the simple division that we have used here. In future work we intend to investigate alternative normalization schemes.

References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* New York: Chapman and Hall.

Edgington, E. S. (1986). Randomization tests. In *Encyclopedia of Statistical Sciences, vol. 7*, eds. Kotz, S., Johnson, N. L., and Read, C. B. New York: Wiley.

Hong, S. J. (1994). Use of contextual information for feature ranking and discretization. *Research Report RC19664*, IBM Research Division, Yorktown Heights, N.Y. To appear in IEEE/TKDE.

Koch, G. G. (1982). Chi-square tests. In *Encyclopedia of Statistical Sciences, vol. 1*, eds. Kotz, S., Johnson, N. L., and Read, C. B. New York: Wiley.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, Calif.: Morgan Kaufmann.

Stuart, A., and Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics, vol. 1*, 5th ed. New York: Oxford University Press.

White, A. P., and Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, **15**, 321–329.