Inf Syst Front (2006) 8:195–209
DOI 10.1007/s10796-006-8779-8

# Splitting methods for decision tree induction: An exploration of the relative performance of two entropy-based families

**Kweku-Muata Osei-Bryson · Kendall Giles**

Received: 4 June 2004 / Revised: 18 May 2005 / Accepted: 17 August 2005

**Abstract** Decision tree (DT) induction is among the more popular of the data mining techniques. An important component of DT induction algorithms is the splitting method, with the most commonly used method being based on the Conditional Entropy (CE) family. However, it is well known that there is no single splitting method that will give the best performance for all problem instances. In this paper we explore the relative performance of the Conditional Entropy family and another family that is based on the Class-Attribute Mutual Information (CAMI) measure. Our results suggest that while some datasets are insensitive to the choice of splitting methods, other datasets are very sensitive to the choice of splitting methods. For example, some of the CAMI family methods may be more appropriate than the popular Gain Ratio (GR) method for datasets which have nominal predictor attributes, and are competitive with the GR method for those datasets where all predictor attributes are numeric. Given that it is never known beforehand which splitting method will lead to the best DT for a given dataset, and given the relatively good performance of the CAMI methods, it seems appropriate to suggest that splitting methods from the CAMI family should be included in data mining toolsets.

**Keywords** Decision trees · Entropy · Splitting methods · Classification · Machine learning

K.-M. Osei-Bryson (✉) · K. Giles
The Information Systems Research Institute, Virginia Commonwealth University, Richmond, VA 23284, U.S.A.
e-mail: Kweku.Muata@isy.vcu.edu

K. Giles
e-mail: KGiles@acm.org

## 1. Introduction

Over the past two decades there has been an increased interest in the use of data mining techniques to address problems in various fields. Among the more popular of these tasks is classification, and for this task various classification algorithms have been proposed, such as decision trees, neural networks, linear discriminant analysis, nonparametric methods, and statistical methods (e.g. Bradley et al., 1999; Wu and Urpani, 1999; Cheesean and Stutz, 1996; Ching et al., 1995; Safavian and Landgrebe, 1991; Quinlan, 1986). In this study we concentrate on decision tree (DT) induction algorithms, and in particular those that use entropy-based splitting methods.

A splitting method is the component of the DT induction algorithm that determines both the attribute that is selected for a given node of the DT and also the partitioning of the values of the selected attribute into mutually exclusive subsets such that each subset uniquely applies to one of the branches that emanate from the given node. Various splitting methods have been proposed (e.g. Breiman, 1984; Lopez de Mantaras, 1991; Martin, 1997; Quinlan, 1986; Quinlan, 1993; Shih, 1999; Taylor, 1993). While the most commonly used splitting methods are based on the Conditional Entropy (CE) family (e.g. Quinlan's C4.5 family of decision tree induction algorithms), it is well known that there is no single splitting method that will give the best performance for all datasets. A question of interest is how well would other families of entropy-based splitting methods measures perform compared to the CE family of entropy-based splitting methods. With this in mind we chose to compare the performance of the Conditional Entropy family and another entropy-based family that is based on a measure called Class-Attribute Mutual Information (CAMI) that was proposed by Ching et al. (1995). While the CE family is well known, the

CAMI family is not as well known. Bryson (2000) identified some of the conceptual links between both families, and developed a new CAMI family splitting method, EffCAMI. In this work we also propose a new hybrid method, adjGR, that involves approaches from both families. Our computational exploration has the objective of testing these splitting methods using a wide variety of datasets from different problem domains and with different data characteristics, and of directly comparing the classification accuracies from both families. Our findings suggest strategies for the data miner to employ when investigating different types of datasets with different splitting measures.

This paper is organized as follows: in Section 2 we present an overview on the entropy-based splitting methods as families. In Section 3 we present the results of our experiments that compare six entropy measures using thirty-five (35) datasets and provide an analysis of these results. Section 4 presents our conclusions.

## 2. Overview on the two entropy-based families

### 2.1. Notations

Let $S$ be the set of classes, $n$ be the total number of examples in the dataset; $n_{j\bullet}$ be the total number of examples in interval $j$ of a given attribute; $n_{\bullet s}$ be the total number of examples in class $s$; $n_{j\cap s}$ be the total number of examples in interval $j$ and class $s$. Also let $p_{j\bullet} = (n_{j\bullet}/n)$ be the estimated probability of being in interval $j$; $p_{\bullet s} = (n_{\bullet s}/n)$ be the estimated probability of being in class $s$; $p_{j\cap s} = (n_{j\cap s}/n)$ be the estimated probability of being in interval $j$ and class $s$; and $p_{s|j} = (n_{j\cap s}/n_{j\bullet}) = (p_{j\cap s}/p_{j\bullet})$ be the conditional probability of an example being in class $s$ given that it is in interval $j$.

### 2.2. Conditional entropy family

The **Information Gain** measure, (e.g. Quinlan, 1993) is based on maximizing the "gain" in information that results from selecting a particular attribute for branching when creating a decision tree. The *Information Gain* (*IG*) for a discretization $\Gamma_g$ of an attribute into $g$ intervals is defined as:

$$IG(g) = -\sum_{s\in S} p_{\bullet s}\log_2(p_{\bullet s}) - \sum_{j\in J(\Gamma g)} p_{j\bullet}\left(-\sum_{s\in S} p_{s|j}\log_2(p_{s|j})\right)$$

where $J(\Gamma_g)$ is the index set of the intervals that are included in a particular discretization $\Gamma_g$ that consists of $g = |\Gamma_g|$ intervals. An examination of the *IG* formula shows a component (i.e., the Unconditional or A Priori Entropy) that is the same for all attributes and all values of $g$, and a second major component that is dependent on the relevant attribute and also on the value of $g$. This second component is called the Conditional Entropy (CE) and is defined as:

$$CE(g) = -\sum_{j\in J(\Gamma g)} p_{j\bullet}\sum_{s\in S} p_{s|j}\log_2(p_{s|j}).$$

An alternate approach to IG involves the maximization of the *Gain Ratio* $GR(g) = \mathrm{IG}(g)/\mathrm{SI}(g)$, where $SI(g) = \sum_{j\in J(\Gamma g)} -p_{j\bullet}\log_2(p_{j\bullet})$, and is called the *Split Information* for the partition $\Gamma_g$ with $g$ intervals by Quinlan (1993). Quinlan (1993) observed that for some datasets the GR measure overcompensates for the bias of IG for higher cardinality attributes and so must be moderated by choosing the attribute with the maximum GR and an above average IG.

### 2.3. Class-Attribute Mutual Information (CAMI) entropy family

The *Class-Attribute Mutual Information (CAMI)* measure proposed by Ching et al. (1995) is a non-decreasing function of the number of intervals $g$, where $g > 1$. The CAMI equation is defined as follows:

$$\mathrm{CAMI}(g) = \sum_{j\in J(\Gamma g)}\sum_{s\in S} p_{j\cap s}\log_2(p_{j\cap s}/p_{j\bullet}p_{\bullet s}),$$

It can be shown that for $g \le |S|$, SupCAMI($g$), the maximum possible value of CAMI($g$), is equal to $\sum_{j\in J(\Gamma g)} -p_{j\bullet}\log_2(p_{j\bullet})$ and is the same as Quinlan's Split Information Measure (Quinlan, 1993) for a discretization of the attribute into $g$ intervals based on partition $\Gamma_g$; for $g \ge |S|$, SupCAMI($g$) is equal to $\sum_{s\in S} -p_{\bullet s}\log_2(p_{\bullet s})$.

Ching et al. (1995) defined a second measure, the Class-Attribute Interdependence Redundancy (CAIR) such that $\mathrm{CAIR}(g) = \mathrm{CAMI}(g)/\mathrm{Max}\{\log_2(|S|), \log_2(g)\}$, and proposed that the attribute discretization problem could be solved by finding the value of $g$ that maximized CAIR($g$). They based this approach on a claim that $\mathrm{Max}\{\log_2(|S|), \log_2(g)\}$ was the maximum value of CAMI(g), and so $\mathrm{CAIR}(g) = 1$ if there is perfect attribute/class interdependency, and $\mathrm{CAIR}(g) = 0$ if there is absolutely no attribute/class interdependency. Bryson (2000) showed that Ching et al.'s assertion, that the maximum possible value of CAMI($g$) is equal to $\mathrm{Max}\{\log_2(|S|), \log_2(g)\}$, is not correct. A more plausible rationale for using CAIR($g$) is that it provides a trade-off between the improvement in the class-attribute mutual information and the cost of the number of intervals. Bryson (2000) proposed a new measure for this family:

$$\mathrm{EffCAMI} = \mathrm{Max}\{\mathrm{CAMI}(g)/\mathrm{SupCAMI}(g),\ g = 2, \ldots, g_{\mathrm{prac}}\}$$

where $g_{\mathrm{prac}}$ is the maximum number of intervals that are appropriate for the given decision tree induction algorithm. Bryson (2000) suggested that EffCAMI could be considered as a measure of the relative strength of the class-attribute interdependence of the given attribute.

## 3. Experimental exploration

### 3.1. Software environment

As mentioned previously we wanted to explore the relative performance of splitting methods that are based on the five entropy measures from both families. We developed software for these splitting methods (i.e., IG, CAMI, CAIR, EffCAMI) based on the Weka library implementation of the well-known C4.5 algorithm. This library (http://www.cs.waikato.ac.nz/~ml/index.html) already contained an implementation of GR, and also provides facilities for pruning, 10-fold cross validation, and calculations. In order to implement the CAMI, CAIR, and EffCAMI splitting methods, we wrote our own Java programs and classes to use the C4.5 algorithm structure in the Weka Java library. Table 1 contains a summary of the entropy measures we used and the decision criteria for each measure. It should be noted that we actually implemented two versions of the EffCAMI measure, as described in Table 1, labeled as EffCAMI_0 and EffCAMI_1. In addition, while the splitting methods should involve the best discretization of each attribute, similar to C4.5 and other DT software, our implementation of these algorithms only involves binarization for continuous attributes, which might not result in the best discretization.

### 3.2. Test problems

We applied our entropy-based splitting methods on 35 publicly available benchmark data mining datasets that we obtained from the UCI Irvine machine library (Murphy and Aha, 1994). These datasets are from a variety of problem domains and have different combinations of nominal and numerical attribute values. Some have missing values and noisy data. The dataset characteristics are summarized in Table 2.

### 3.3. Test results—Performance of two families of entropy measures

While there are several criteria used to judge optimal tree design (e.g. Safavian and Landgrebe, 1991), we used the most commonly used measure, classification accuracy rate, as the indicator of algorithm performance. It should be noted that these classification accuracy rates are based on the application of stratified ten-fold cross-validation.

**Table 1** Induction algorithm decision rules for selecting the best attribute

| Entropy measure | Decision rule |
|---|---|
| GR | For those attributes whose IG > Average(IG), select the attribute that provides Max(GR). |
| IG | Select the attribute that provides Max(IG) |
| CAMI | Select the attribute that provides Max(CAMI) |
| CAIR | For those attributes whose CAMI > Average(CAMI), select the attribute that provides Max(CAIR) |
| EffCAMI_0 | For those attributes whose CAMI > Average(CAMI), select the attribute that provides Max(EffCAMI) |
| EffCAMI_1 | For those attributes whose CAMI > Average(CAMI)—StandardDeviation (CAMI), select the attribute that provides Max(EffCAMI) |

The reader may observe (see Table 3) that while each entropy measure has varying results over all the datasets, it is interesting to note that no one particular entropy measure stands out above all the others *over the collection of datasets* used in the study. However, a closer examination of these results reveals that for the some datasets (e.g. *Page Blocks, Mushroom, Chess, Letter, Segment, Sick*) there is only a marginal difference in the performance of the splitting methods, while for other datasets the difference in the performance is more obvious, and for a few of the problems the differences are large (e.g. *Audiology, Colic, Labor, Glass, Soybean, Heart*). For example, with the *Audiology* and *Labor* datasets the largest difference was over 10%, and with the *Heart* dataset the largest difference was over 5%. In some cases, such as with the *Soybean* dataset, the within-family differences were greater than the differences in the best performances of both families. In other cases, such as with the *Heart* dataset, one family outperformed the other. This led us to investigate further the differences in classification accuracies between the two families (see Tables 4(a) and (b)).

For each dataset, we compared the accuracy rate of the best method from each family using two approaches. The first approach involved a statistical difference of proportions test at the $\alpha = 0.05$ level of significance (see Statistical column of Table 4(a)). The second approach involved the examination of the arithmetic difference between both accuracy rates, where differences that were less than the tolerance $\tau = 0.5\%$ were considered to be a tie (see Arithmetic column of Table 4(a)).

We will first consider the results of the statistical difference of proportions tests. The value "**NONE**" in the Statistical column of Table 4(a) indicates that the difference between the

**Table 2** Dataset characteristics

| Dataset | | | | Predictor attributes | |
|---|---|---|---|---|---|
| ID | Name | # Instances | # Attributes | Domain group | Data type(s) |
| 1 | IRIS | 150 | 5 | Ordered | C |
| 2 | Breast Cancer | 349 | 10 | Ordered | I |
| 3 | Credit Approval | 690 | 16 | Mixed | N, C, I |
| 4 | Car | 1728 | 7 | Nominal | N |
| 5 | Abalone | 4177 | 9 | Mixed | N, C |
| 6 | Wave | 5000 | 22 | Ordered | C |
| 7 | Glass | 214 | 10 | Ordered | C |
| 8 | Soybean | 683 | 36 | Nominal | N |
| 9 | Page Blocks | 5473 | 11 | Ordered | C, I |
| 10 | Mushroom | 8124 | 23 | Nominal | N |
| 11 | Wine | 178 | 14 | Ordered | C |
| 12 | Yeast | 1484 | 9 | Ordered | C |
| 13 | Zoo | 101 | 17 | Mixed | N, I |
| 14 | Pima | 768 | 9 | Ordered | C, I |
| 15 | Nursery | 12960 | 9 | Nominal | N |
| 16 | Audiology | 226 | 70 | Nominal | N |
| 17 | Heart | 270 | 14 | Ordered | C, I |
| 18 | Hepatitis | 155 | 20 | Mixed | N, C, I |
| 19 | Tumor | 339 | 18 | Nominal | N |
| 20 | Chess | 3196 | 37 | Nominal | N |
| 21 | Letter | 20000 | 17 | Ordered | I |
| 22 | Segment | 2310 | 20 | Ordered | C, I |
| 23 | Sick | 3772 | 30 | Ordered | C, I |
| 24 | Sonar | 208 | 61 | Ordered | C |
| 25 | Splice | 3190 | 61 | Nominal | N |
| 26 | Anneal | 898 | 39 | Mixed | N, C, I |
| 27 | Autos | 205 | 26 | Mixed | N, C, I |
| 28 | Colic | 368 | 23 | Mixed | N, C, I |
| 29 | Hypothyroid | 3772 | 30 | Mixed | N, C, I |
| 30 | Ionosphere | 351 | 35 | Ordered | C,I |
| 31 | Labor | 57 | 17 | Mixed | N, C, I |
| 32 | Lymph | 148 | 19 | Mixed | N, I |
| 33 | Vehicle | 846 | 19 | Ordered | I |
| 34 | Vote | 435 | 17 | Nominal | N |
| 35 | Vowel | 990 | 14 | Mixed | N, C |

**C:** Continuous, **I:** Integer, **N:** Norminal

relevant pair of accuracy rates is not considered to be statistically significant at the $\alpha = 0.05$ level of significance. In the "CI Width" column Table 4(a) displays the width of the confidence interval associated with this statistical test, and the "BestCE–BestCAMI" column Table 4(a) displays the difference between the relevant pair of accuracy rates. The reader may observe that for the *Glass*, *Heart* and *Labor* datasets the corresponding differences (i.e., 4.20%, 4.45%, 8.77%) are not considered to be statistically significant, but for the *Nursery* dataset the corresponding difference (i.e., 1.02%) is considered to be statistically significant. Thus, using our statistical difference of proportions test, the difference between the best accuracy rates of the two families is considered to be statistically significant for only one of the datasets (i.e., *Nursery*). Thus using this test the performance of the best splitting method of the CAMI family would be considered to be no worse and marginally better than the corresponding performance of the best splitting method of the CE family.

It should be noted that for several other datasets, apart from *Nursery*, that there are pairs of splitting methods for which the respective difference in the accuracy rates is statistically significant. For each such pair, at least one of the splitting methods does give the best accuracy rate for the relevant family. Table 4(b) displays some of these results. The reader may observe that for the *Vowel* dataset, the difference between GR and CAIR is statistically significant in favor of CAIR. Similarly for the *Colic* dataset, the difference between GR and CAIR is statistically significant in favor of CAIR, and the difference between GR and IG is statistically significant in favor of IG. The latter result is particularly interesting since GR and IG are members of the same family, and GR is normally considered to be superior to IG. Also for the *Wave*

**Table 3** Classification accuracy of CE and CAMI families

| Dataset | | CE family | | CAMI family | | | | |
|---|---|---|---|---|---|---|---|---|
| ID | Name | GR | IG | CAMI | CAIR | EffCAMI_0 | EffCAMI_1 | Best-Worst |
| 1 | IRIS | **95.33** | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 0.66 |
| 2 | Breast cancer | 72.49 | 72.49 | 72.49 | 72.49 | **74.50** | **74.50** | 2.01 |
| 3 | Credit approval | 85.94 | 84.20 | 85.36 | **86.96** | 84.64 | 84.64 | 2.76 |
| 4 | Car | 92.48 | **93.52** | **93.52** | **93.52** | 92.48 | 92.48 | 1.04 |
| 5 | Abalone | 20.19 | 20.79 | 20.76 | 20.71 | 20.40 | 21.88 | 1.69 |
| 6 | Wave | 77.02 | 76.74 | 76.76 | 76.76 | 75.66 | 75.24 | 1.78 |
| 7 | Glass | 65.89 | 67.76 | 67.76 | 67.76 | **71.96** | 69.63 | 6.07 |
| 8 | Soybean | 92.09 | 87.56 | 89.02 | 89.02 | 92.09 | **92.68** | 5.12 |
| 9 | Page blocks | 96.95 | 96.99 | 96.99 | 96.99 | 96.97 | 96.78 | 0.21 |
| 10 | Mushroom | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.00 |
| 11 | Wine | 94.94 | **95.51** | **95.51** | **95.51** | 94.94 | 94.94 | 0.57 |
| 12 | Yeast | 54.78 | 53.03 | 53.03 | 53.03 | **55.39** | 54.72 | 2.36 |
| 13 | Zoo | 92.08 | **94.06** | **94.06** | **94.06** | 92.08 | 92.08 | 1.98 |
| 14 | Pima | **74.09** | 72.14 | 72.14 | 72.14 | 71.48 | 71.61 | 2.61 |
| 15 | Nursery | 97.11 | 97.10 | 97.10 | **98.13** | 97.11 | 97.11 | 1.03 |
| 16 | Audiology | 77.88 | 67.26 | 77.43 | 77.43 | 77.88 | **79.65** | 12.39 |
| 17 | Heart | **77.78** | 72.59 | 72.59 | 72.59 | 73.33 | 73.33 | 5.19 |
| 18 | Hepatitis | 79.35 | 78.06 | **80.65** | **80.65** | 80.00 | 80.00 | 2.59 |
| 19 | Tumor | 40.71 | 43.01 | 40.41 | 40.41 | 42.18 | **44.25** | 3.84 |
| 20 | Chess | 99.53 | 99.44 | 99.44 | 99.44 | 99.47 | 99.47 | 0.09 |
| 21 | Letter | 87.76 | 87.96 | 87.96 | 87.96 | 87.68 | 87.67 | 0.29 |
| 22 | Segment | 97.14 | 97.10 | 97.10 | 97.10 | 96.93 | 96.97 | 0.21 |
| 23 | Sick | 98.65 | 98.52 | 98.97 | 98.91 | 98.67 | 98.67 | 0.45 |
| 24 | Sonar | 74.04 | 73.08 | 73.08 | 73.08 | **75.97** | **75.96** | 2.89 |
| 25 | Splice | 93.98 | 93.67 | 93.67 | *93.51* | 93.67 | 93.67 | 0.47 |
| 26 | Anneal | 98.44 | 98.89 | 98.89 | 98.55 | 98.55 | 98.55 | 0.45 |
| 27 | Autos | **82.44** | 73.17 | 77.56 | 77.56 | 80.00 | 79.02 | 9.27 |
| 28 | Colic | **85.87** | 76.09 | **85.60** | **85.87** | **85.87** | **85.87** | 9.78 |
| 29 | Hypothyroid | **99.58** | 98.99 | **99.50** | **99.50** | **99.55** | **99.55** | 0.59 |
| 30 | Ionosphere | **90.88** | 88.60 | 88.60 | 88.60 | 90.03 | 89.46 | 2.28 |
| 31 | Labor | 75.44 | 71.93 | 80.70 | **84.21** | 78.95 | 78.95 | 12.28 |
| 32 | Lymph | **77.03** | **77.03** | **77.03** | 76.35 | 75.00 | 75.68 | 2.03 |
| 33 | Vehicle | 73.40 | 73.17 | 73.17 | 73.17 | 73.52 | 73.40 | 0.35 |
| 34 | Vote | **96.78** | 95.86 | **96.55** | **96.55** | **96.55** | **96.55** | 0.92 |
| 35 | Vowel | 78.38 | 83.84 | 83.84 | **85.45** | 79.49 | 78.59 | 7.07 |

dataset the difference between GR and EffCAMI_1 is statistically significant in favor or GR, while for the *Soybean* dataset the difference between IG and EffCAMI_1 is statistically significant in favor or EffCAMI_1. This analysis demonstrates that no splitting method is dominant for all datasets, and for some datasets the differences in the performances of certain pairs of splitting methods are statistically significant, thus demonstrating that some datasets are sensitive to the choice of splitting methods.

The second approach to comparing the performances of the splitting methods involved the examination of the arithmetic difference between the accuracy rates of each pair of splitting methods, where differences that were less than the tolerance $\tau = 0.5\%$ were considered to be a tie (see Arithmetic column of Table 4(a)). It should be noted that even when the difference in the classification rates are not statistically significant, a decision still has to be made as to which DT should be used. Thus, many post-pruning methods (including those used in commercial data mining software such as SAS Enterprise Miner) also use the best accuracy rate on the validation dataset to select the best sub-tree even if the difference is not statistically significant. Some of these commercial DM software round to the third decimal position when classification accuracy is represented as a proportion, which is equivalent to using a tolerance of $\tau = 0.5\%$. For the remainder of this subsection all relative performance comparisons are based on the arithmetic difference.

An examination of the Arithmetic column of Table 4(a) shows that the CE family outperforms the CAMI family five (5) times, the CAMI family outperforms the CE family thirteen (13) times, and there are seventeen (17) ties. These results suggest that although the splitting methods from the CE

**Table 4(a)** Difference in classification accuracy between best of CE and CAMI families

| | | | | | | BestCE - | Winner | |
|---|---|---|---|---|---|---|---|---|
| ID | Name | BestCE | BestCAMI | N | CI Width | BestCAMI | Statistical | Arithmetic |
| 1 | IRIS | **95.33** | 94.67 | 150 | 4.93 | 0.66 | **NONE** | CE |
| 2 | Breast Cancer | 72.49 | **74.5** | 349 | 6.55 | −2.01 | **NONE** | CAMI |
| 3 | Credit Approval | 85.94 | **86.96** | 690 | 3.61 | −1.02 | **NONE** | CAMI |
| 4 | Car | 93.52 | 93.52 | 1728 | 1.64 | 0.00 | **NONE** | **TIE** |
| 5 | Abalone | 20.79 | **21.88** | 4177 | 1.76 | −1.09 | **NONE** | CAMI |
| 6 | Wave | 77.02 | 76.76 | 5000 | 1.65 | 0.26 | **NONE** | **TIE** |
| 7 | Glass | 67.76 | **71.96** | 214 | 8.69 | −4.20 | **NONE** | CAMI |
| 8 | Soybean | 92.09 | **92.68** | 683 | 2.81 | −0.59 | **NONE** | CAMI |
| 9 | Page Blocks | 96.99 | 96.99 | 5473 | 0.64 | 0.00 | **NONE** | **TIE** |
| 10 | Mushroom | 100 | 100 | 8124 | 0.00 | 0.00 | **NONE** | **TIE** |
| 11 | Wine | 95.51 | 95.51 | 178 | 4.30 | 0.00 | **NONE** | **TIE** |
| 12 | Yeast | 54.78 | **55.39** | 1484 | 3.58 | −0.61 | **NONE** | CAMI |
| 13 | Zoo | 94.06 | 94.06 | 101 | 6.52 | 0.00 | **NONE** | **TIE** |
| 14 | Pima | **74.09** | 72.14 | 768 | 4.43 | 1.95 | **NONE** | CE |
| 15 | Nursery | 97.11 | **98.13** | 12960 | 0.37 | −1.02 | CAMI | CAMI |
| 16 | Audiology | 77.88 | **79.65** | 226 | 7.54 | −1.77 | **NONE** | CAMI |
| 17 | Heart | **77.78** | 73.33 | 270 | 7.24 | 4.45 | **NONE** | CE |
| 18 | Hepatitis | 79.35 | **80.65** | 155 | 8.90 | −1.30 | **NONE** | CAMI |
| 19 | Tumor | 43.01 | **44.25** | 339 | 7.47 | −1.24 | **NONE** | CAMI |
| 20 | Chess | *99.53* | 99.47 | 3196 | 0.35 | 0.06 | **NONE** | **TIE** |
| 21 | Letter | 87.96 | 87.96 | 20000 | 0.64 | 0.00 | **NONE** | **TIE** |
| 22 | Segment | 97.14 | 97.1 | 2310 | 0.96 | 0.04 | **NONE** | **TIE** |
| 23 | Sick | 98.65 | *98.97* | 3772 | 0.49 | −0.32 | **NONE** | **TIE** |
| 24 | Sonar | 74.04 | **75.97** | 208 | 8.32 | −1.93 | **NONE** | CAMI |
| 25 | Splice | *93.98* | 93.67 | 3190 | 1.18 | 0.31 | **NONE** | **TIE** |
| 26 | Anneal | *98.89* | 98.89 | 898 | 0.97 | 0.00 | **NONE** | **TIE** |
| 27 | Autos | ***82.44*** | 80 | 205 | 7.56 | 2.44 | **NONE** | CE |
| 28 | Colic | *85.87* | 85.87 | 368 | 5.03 | 0.00 | **NONE** | **TIE** |
| 29 | Hypothyroid | *99.58* | 99.55 | 3772 | 0.30 | 0.03 | **NONE** | **TIE** |
| 30 | Ionosphere | ***90.88*** | 90.03 | 351 | 4.35 | 0.85 | **NONE** | CE |
| 31 | Labor | *75.44* | **84.21** | 57 | 14.65 | −8.77 | **NONE** | CAMI |
| 32 | Lymph | *77.03* | 77.03 | 148 | 9.58 | 0.00 | **NONE** | **TIE** |
| 33 | Vehicle | *73.4* | 73.52 | 846 | 4.21 | −0.12 | **NONE** | **TIE** |
| 34 | Vote | *96.78* | 96.55 | 435 | 2.39 | 0.23 | **NONE** | **TIE** |
| 35 | Vowel | *83.84* | **85.45** | 990 | 3.18 | −1.61 | **NONE** | CAMI |

****NONE**** indicates difference is not statistically significant at significance level $\alpha = 0.05$.
****TIE**** indicates that the Arithmetic difference in the Classification Accuracy is 0.50 or less.

**Table 4(b)** Some performance differences that are statistically significant

| ID | Name | Test | Difference | Statistically significant |
|---|---|---|---|---|
| 6 | Wave | GR > EffCAMI_1 | 1.78 | YES |
| 8 | Soybean | EffCAMI_1 > IG | 5.12 | YES |
| 15 | Nursery | CAIR > GR | 1.02 | YES |
| 16 | Audiology | EffCAMI_0 > IG | 10.62 | YES |
| 28 | Colic | EffCAMI_0 > IG | 9.78 | YES |
| 35 | Vowel | CAIR > GR | 7.07 | YES |
| | | IG > GR | 5.46 | YES |

family are the more commonly used methods in DT induction algorithms, it might be worthwhile to do further exploration of various methods from the CAMI families. For example, we then thought to explore the relative performance of EffCAMI_0 and EffCAMI_1, our new methods from the CAMI family, versus that of the commonly used GR method (see Table 5).

The information in Table 5 suggests that based on the arithmetic differences in the accuracy rates: (a) CAIR performs better than GR for the Mixed domain group datasets; (b) EffCAMI_1 and CAIR may perform better than GR for

**Table 5** Selected pairwise comparisons of splitting methods

| ID | Domain group | Data types | GR: IG | GR: EffCAMI_0 | GR: EffCAMI_1 | GR: CAIR | CAIR: EffCAMI_0 | CAIR: EffCAMI_1 | CAMI: CAIR |
|---|---|---|---|---|---|---|---|---|---|
| 5 | Mixed | N, C | IG | **TIE** | EffCAMI_1 | **TIE** | **TIE** | EffCAMI_1 | **TIE** |
| 35 | Mixed | N, C | IG | EffCAMI_0 | **TIE** | CAIR | CAIR | CAIR | CAIR |
| 3 | Mixed | N, C, I | GR | GR | GR | CAIR | CAIR | CAIR | CAIR |
| 18 | Mixed | N, C, I | GR | EffCAMI_0 | EffCAMI_1 | CAIR | CAIR | CAIR | **TIE** |
| 26 | Mixed | N, C, I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 27 | Mixed | N, C, I | GR | GR | GR | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 28 | Mixed | N, C, I | GR | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 29 | Mixed | N, C, I | GR | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 31 | Mixed | N, C, I | GR | EffCAMI_0 | EffCAMI_1 | CAIR | CAIR | CAIR | CAIR |
| 13 | Mixed | N, I | IG | **TIE** | **TIE** | CAIR | CAIR | CAIR | **TIE** |
| 32 | Mixed | N, I | **TIE** | GR | GR | GR | CAIR | CAIR | CAMI |
| 4 | Nominal | N | IG | **TIE** | **TIE** | CAIR | CAIR | CAIR | **TIE** |
| 8 | Nominal | N | GR | **TIE** | EffCAMI_1 | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 10 | Nominal | N | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 15 | Nominal | N | **TIE** | **TIE** | **TIE** | CAIR | CAIR | CAIR | CAIR |
| 16 | Nominal | N | GR | **TIE** | EffCAMI_1 | **TIE** | **TIE** | EffCAMI_1 | **TIE** |
| 19 | Nominal | N | IG | EffCAMI_0 | EffCAMI_1 | **TIE** | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 20 | Nominal | N | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 25 | Nominal | N | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 34 | Nominal | N | GR | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 1 | Ordered | C | GR | GR | GR | GR | **TIE** | **TIE** | **TIE** |
| 6 | Ordered | C | **TIE** | GR | GR | CAIR | CAIR | CAIR | **TIE** |
| 7 | Ordered | C | IG | EffCAMI_0 | EffCAMI_1 | **TIE** | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 11 | Ordered | C | IG | **TIE** | **TIE** | CAIR | CAIR | CAIR | **TIE** |
| 12 | Ordered | C | GR | EffCAMI_0 | **TIE** | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 24 | Ordered | C | GR | EffCAMI_0 | EffCAMI_1 | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 9 | Ordered | C, I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 14 | Ordered | C, I | GR | GR | GR | GR | CAIR | CAIR | **TIE** |
| 17 | Ordered | C, I | GR | GR | GR | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 22 | Ordered | C, I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 23 | Ordered | C, I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 30 | Ordered | C, I | GR | GR | GR | GR | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 2 | Ordered | I | **TIE** | EffCAMI_0 | EffCAMI_1 | **TIE** | EffCAMI_0 | EffCAMI_1 | **TIE** |
| 21 | Ordered | I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |
| 33 | Ordered | I | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** | **TIE** |

**C:** Continious **I:** Integer **N:** Nominal

Nominal domain group datasets; and (c) GR and EffCAMI_1 performances are approximately the same for the Ordered domain group datasets (i.e., all predictor attributes are either continuous or integer). It appears that overall the CAMI family does better than the CE family if the domain group is Nominal or Mixed, while the performances of both families are approximately the same whenever the domain group is Ordered. In fact, CAIR appears to be a better choice than GR for the Mixed domain group datasets while EffCAMI_1 appears to be a better choice than GR for the Nominal domain group datasets.

This led us to consider whether a new CE family measure would perform better than GR when the domains are Nominal or Mixed. The reader may recall that $\mathrm{GR}(g) = \mathrm{IG}(g)/\sum_{j\in J\Gamma g} -p_{j\bullet}\log_2(p_{j\bullet})$, if the predictor attribute takes it values from Nominal or Ordered domain. The situation with EffCAMI varies with the domain of the predictor attribute. If the number of intervals is not greater than the number of Classes (i.e., $g \leq |S|$), which is always the case for the variables from the Ordered domain since we are only doing binary cuts, then $\mathrm{EffCAMI} = \mathrm{CAMI}(g)/\sum_{j\in J\Gamma g} -p_{j\bullet}\log_2(p_{j\bullet})$. On the other hand if the number of intervals is greater than the number of Classes (i.e., $g > |S|$) then $\mathrm{EffCAMI} = \mathrm{CAMI}(g)/\sum_{s\in S} -p_{\bullet s}\log_2(p_{\bullet s})$. CAIR is similar to EffCAMI in this regard.

We, therefore, developed a new measure adjGR that is the same as GR for integer and continuous data variables when binarization is used, but different for Nominal variables:

- $\mathrm{adjGR}(g) = \mathrm{IG}(g)/\sum_{j\in J\Gamma g} -p_{j\bullet}\log_2(p_{j\bullet})$ if $g \leq |S|$
- $\mathrm{adjGR}(g) = \mathrm{IG}(g)/\sum_{s\in S} -p_{\bullet s}\log_2(p_{\bullet s})$ if $g > |S|$.

**Table 6** Comparison of GainRatio(GR) and adjGR methods

| ID | Dataset | Domain group | Data types | GR | adjGR | \|GR-adjGR\| |
|---|---|---|---|---|---|---|
| 31 | Labor | Mixed | N, C, I | 75.44 | **80.70** | 5.26 |
| 35 | Vowel | Mixed | N, C | 78.38 | **83.03** | 4.65 |
| 27 | Autos | Mixed | N, C, I | **82.44** | 78.05 | 4.39 |
| 3 | Credit Approval | Mixed | N, C, I | **85.94** | 85.07 | 0.87 |
| 32 | Lymph | Mixed | N, I | 77.03 | **77.70** | 0.67 |
| 26 | Anneal | Mixed | N, C, I | 98.44 | 98.55 | 0.11 |
| 5 | Abalone | Mixed | N, C | 20.19 | 20.18 | 0.01 |
| 13 | Zoo | Mixed | N, I | 92.08 | 92.08 | 0.00 |
| 18 | Hepatitis | Mixed | N, C, I | 79.35 | 79.35 | 0.00 |
| 28 | Colic | Mixed | N, C, I | 85.87 | 85.87 | 0.00 |
| 29 | Hypothyroid | Mixed | N, C, I | 99.58 | 99.58 | 0.00 |
| 25 | Splice | Nominal | N | 93.98 | 93.67 | 0.31 |
| 4 | Car | Nominal | N | 92.48 | 92.48 | 0.00 |
| 8 | Soybean | Nominal | N | 92.09 | 92.09 | 0.00 |
| 10 | Mushroom | Nominal | N | 100.00 | 100.00 | 0.00 |
| 15 | Nursery | Nominal | N | 97.11 | 97.11 | 0.00 |
| 16 | Audiology | Nominal | N | 77.88 | 77.88 | 0.00 |
| 19 | Tumor | Nominal | N | 40.71 | 40.71 | 0.00 |
| 20 | Chess | Nominal | N | 99.53 | 99.53 | 0.00 |
| 34 | Vote | Nominal | N | 96.78 | 96.78 | 0.00 |
| 23 | Sick | Ordered | C, I | 98.65 | 98.70 | 0.05 |
| 1 | IRIS | Ordered | C | 95.33 | 95.33 | 0.00 |
| 2 | Breast Cancer | Ordered | I | 72.49 | 72.49 | 0.00 |
| 6 | Wave | Ordered | C | 77.02 | 77.02 | 0.00 |
| 7 | Glass | Ordered | C | 65.89 | 65.89 | 0.00 |
| 9 | Page Blocks | Ordered | C, I | 96.95 | 96.95 | 0.00 |
| 11 | Wine | Ordered | C | 94.94 | 94.94 | 0.00 |
| 12 | Yeast | Ordered | C | 54.78 | 54.78 | 0.00 |
| 14 | Pima | Ordered | C, I | 74.09 | 74.09 | 0.00 |
| 17 | Heart | Ordered | C, I | 77.78 | 77.78 | 0.00 |
| 21 | Letter | Ordered | I | 87.76 | 87.76 | 0.00 |
| 22 | Segment | Ordered | C, I | 97.14 | 97.14 | 0.00 |
| 24 | Sonar | Ordered | C | 74.04 | 74.04 | 0.00 |
| 30 | Ionosphere | Ordered | C, I | 90.88 | 90.88 | 0.00 |
| 33 | Vehicle | Ordered | I | 73.40 | 73.40 | 0.00 |

Since for Ordered domain group attributes, adjGR is the same as GR, then for this domain group the performance of adjGR should be the same as that of GR. For Nominal group attributes, given the relatively poor performance of the IG method, we decided to use the CAMI method for selecting the 'best' split for each appropriate nominal attribute and then to calculate the corresponding adjGR value for the given attribute using the formula above.

Our results obtained from applying this method (see Table 6) show that the adjGR method mirrored GR, except for a few of the Mixed domain group datasets where it beat GR and a few of the Mixed domain group datasets where it was beaten by GR. As expected, adjGR gave the same performance as GR on the Ordered domain datasets. While we had hoped that the relative performance of adjGR to GR would be clearer for Nominal domain group datasets, the results still suggest that: (1) some datasets are sensitive to choice of splitting methods, while others are not; and (2) although GR is currently the popular choice in most data mining software, adjGR is just as good.

### 3.4. Analysis based on tree sizes

We will now compare the tree sizes, in terms of the average number of leaves (based on 10-fold cross validation), produced by the splitting methods for each dataset. For several of the datasets (e.g. *Iris*, *Wine*) there is no arithmetic or statistically significant difference in the performance of the different splitting methods (see Table A1 in Appendix A). For other datasets the differences are significant, both arithmetically and statistically. Table 7 below provides some examples when the differences are statistically significant. In this table we also explore the relative performance in terms of tree size with the corresponding relative performance in terms of accuracy rates.

**Table 7** Some tree size differences that are statistically significant

| | | Number of leaves | | | Accuracy Rates | | |
|---|---|---|---|---|---|---|---|
| ID | Dataset | Hypothesis | Difference | Statistically Significant | Hypothesis | Difference | Statistically Significant |
| 6 | Wave | GR < EffCAMI_1 | −77.80 | YES | GR > EffCAMI_1 | 1.78% | YES |
| 18 | Hepatitis | CAIR < GR | −2.70 | YES | CAIR > GR | 1.30% | NO |
| 25 | Splice | EffCAMI_1 < GR | −14.40 | YES | EffCAMI_1 ≅ GR | −0.31% | NO |
| 32 | Lymph | EffCAMI_0 < GR | −6.90 | YES | EffCAMI_0 > GR | 3.51% | NO |

- For the *Wave* dataset, GR outperforms EffCAMI_1 with regard to both the Number of Leaves (i.e., the hypothesis, tree size based on GR is better than the tree size based on EffCAMI_1, would be accepted) and Accuracy Rate measures (i.e., the hypothesis, accuracy rate based on GR is better than the accuracy rate based on EffCAMI_1, would be accepted). In the language of bi-objective programming, GR would be said to dominate EffCAMI_1 for the *Wave* dataset. Thus for this dataset GR can be considered to have better overall performance than EffCAMI_1 in terms of these two measures.
- For the *Hepatitis* dataset, CAIR outperforms GR with regard to the Number of Leaves (i.e., the hypothesis tree size based on CAIR is better than the tree size based on GR would be accepted). With regard to the Accuracy Rate measure, the hypothesis that accuracy rate based on CAIR is better than the accuracy rate based on GR would not be accepted even though the arithmetic difference is in favor of CAIR. In the language of bi-objective programming, GR would still be said to dominate EffCAMI_1 for the *Wave* dataset since it outperforms it on one measure (i.e., Number of Leaves) and is not inferior to it in terms of the other measure. Thus for this dataset CAIR would be considered to have better overall performance than GR in terms of these two measures.
- Similarly for the *Splice* dataset, EffCAMI_1 would be considered to have better overall performance than GR in terms of these two measures.
- Similarly for the *Lymph* dataset, EffCAMI_0 would be considered to have better overall performance than GR in terms of these two measures.

These results suggest that neither family dominates the other in terms of these two performance measures, and provides further evidence that the CAMI family is competitive to the CE family, and that certain datasets are sensitive to the choice of splitting methods.

### 3.5. Analysis based on unequal misclassification costs

The results in the previous subsections assumed that for each dataset the misclassification costs for the relevant target events (i.e., classes) were the same. In this subsection, we focus on the case when these costs are unequal. Our experimental work involved those of our datasets that have binary target events, and examined the impact of different cost ratios for the relevant target events. Table A2 in Appendix 2 provides a detailed description of the results which implies that no single method is dominant for all datasets, and that for a given dataset the relative performance of a splitting method could vary with the cost ratio. Figure 1 displays this phenomenon graphically. These results also demonstrate that the performances of splitting methods from the CAMI family and the new splitting method adjGR are competitive to that of splitting methods from the CE family.

### 3.6. Discussion

A few observations can be made from the evidence presented in Tables 3 through 7:

1. Some datasets are insensitive to the choice of splitting methods (e.g. *Page Blocks, Mushroom, Chess, Letter, Segment, Sick*) while other datasets are sensitive to the choice of splitting methods, with some being extremely sensitive (e.g. *Audiology, Colic, Glass, Labor, Soybean, Heart*). However, for the most part the data miner is never clear at the start of the data mining project as to whether a given dataset is sensitive or insensitive to the choice of splitting method.
2. As has been established previously, no single splitting method is likely to perform best on all datasets.
3. If the dataset only consists of continuous predictor attributes and the splitting method only does binary discretization, then there is no difference between CAMI and CAIR because for each attribute CAIR $= \text{CAMI}/\log_2(2)$, and so for each node of the DT the choice of attribute based on CAIR would be the same as that based on CAMI.
4. The CAMI family of splitting methods and the new splitting method adjGR performed impressively compared to the more popular CE family.

Given these observations it seems appropriate to suggest that splitting methods from the CAMI family and the adjGR splitting method should be included in data mining toolsets. Although the question might be raised as to which splitting

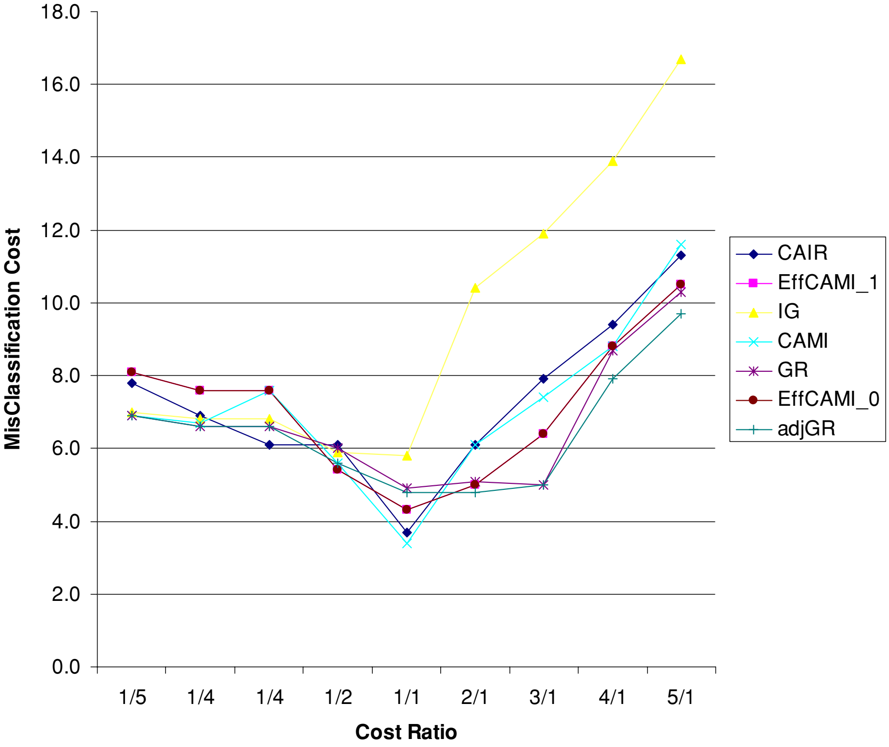

**Fig. 1** Sick dataset—comparison of misclassification costs

method should be selected by the data miner, the fact is that it is never known beforehand which splitting method will lead to the best DT for the given dataset. Many modern data mining tools provide multiple options for splitting methods. For example SAS Enterprise Miner offers the data miner the option of selecting either Chi Squared, Entropy (i.e., Gain Ratio), or Gini splitting methods. Also modern data mining tools typically provide the data miner with multiple parameters (e.g. depth of DT, pre-pruning and post-pruning rules, splitting method) with multiple options for each. Gersten et al. (2000) notes that with regard to setting parameter values, there is "no practicable approach to select . . . the most promising combinations early in the process" and as such "it is necessary to experiment with different combinations" in order to be able to reliably pick the best DT. The process of DT induction in an industrial setting thus involves experimentation with different combinations of parameter settings and in some cases with different training and validation datasets in order to be able to select the most appropriate decision tree. Therefore, given that data miners already experiment with different splitting methods it would be worthwhile to include methods from a family that performs impressively against the currently most popular method.

If the given dataset is relatively small it might be possible to apply several methods and then use the one that gives the best performance with regard to measures such as accuracy, stability, and simplicity. If the given dataset is relatively large, it might be very costly to explore the performance of several splitting methods on the entire dataset. One approach is to take a sample from the given dataset, apply the different splitting methods to this sample, and then apply the splitting method that gave the best performance to the entire dataset. It should be noted that such an approach is also used in industrial applications, and as such some commercial data mining tools provide convenient facilities for sampling of the dataset.

## 4. Conclusions

In this paper we conducted a computational exploration of the performance of the two families of entropy-based splitting methods, Conditional Entropy and Class-Attribute Mutual Information (CAMI). Our results suggest that while some datasets are insensitive to the choice of splitting methods, others are very sensitive to the choice of splitting method. In summary, our results suggest that: (1) some of the CAMI family methods may be more appropriate than the commonly used GR method for datasets where all predictor attributes are nominal; (2) that if the only type of discretization on continuous attributes is binarization then some of the CAMI methods perform as well as GR for datasets where all the predictor attributes are either integer or continuous; and (3) that the new EffCAMI_1 method and the older CAIR method performed very well, particularly when compared to the popular GR method. Given these results it seems appropriate to suggest that splitting methods from the CAMI family should be included in data mining toolsets.

## Appendix 1: Detailed results of tree size analysis

**Table A1** Average number of leaves of pruned DTs

| Dataset | | CE family | | CAMI family | | | | |
|---|---|---|---|---|---|---|---|---|
| ID | Name | GR | IG | CAMI | CAIR | EffCAMI_0 | EffCAMI_1 | Hybrid adjGR |
| 1 | Iris | **4.7** | 4.8 | 4.8 | 4.8 | 4.8 | 4.8 | 4.7 |
| 2 | Breast Cancer | 13.2 | 11.3 | 11.3 | 11.3 | **11.0** | 11.1 | 13.2 |
| 3 | Credit Approval | 22.2 | 28.5 | 32.0 | **21.5** | 33.4 | 33.4 | 30.6 |
| 4 | Car | **122.2** | 125.6 | 125.6 | 125.6 | **122.2** | **122.2** | **122.2** |
| 5 | Abalone | **1051.4** | 1079.2 | 1079.2 | 1079.8 | 1107.3 | 1097.1 | **1051.4** |
| 6 | Wave | **276.1** | 321.8 | 321.8 | 321.8 | 347.9 | 353.9 | **276.1** |
| 7 | Glass | **23.0** | 26.3 | 26.3 | 26.3 | 26.2 | 25.6 | **23.0** |
| 8 | Soybean | 62.5 | **56.8** | 74.3 | 74.3 | 60.1 | 61.7 | 62.5 |
| 9 | Page Blocks | **41.8** | 45.6 | 45.6 | 45.6 | 45.4 | 46.1 | **41.8** |
| 10 | Mushroom | 25.0 | 33.0 | 33.0 | 23.0 | 33.0 | 33.0 | **22.0** |
| 11 | Wine | 5.4 | **5.3** | **5.3** | **5.3** | 5.4 | 5.4 | 5.4 |
| 12 | Yeast | **165.9** | 198.9 | 198.9 | 198.9 | 185.2 | 187.4 | **165.9** |
| 13 | Zoo | **10.7** | 12.7 | 12.7 | 12.7 | **10.7** | **10.7** | **10.7** |
| 14 | Pima | **20.0** | 59.8 | 59.8 | 59.8 | 64.1 | 63.9 | **20.0** |
| 15 | Nursery | **353.1** | 353.2 | 353.2 | 353.2 | **353.1** | **353.1** | **353.1** |
| 16 | Audiology | 30.6 | **11.6** | 34.4 | 34.4 | 30.6 | 31.1 | 30.6 |
| 17 | Heart | **17.4** | 19.5 | 19.5 | 19.5 | 19.2 | 19.2 | **17.4** |
| 18 | Hepatitis | 9.5 | **3.4** | 6.8 | 6.8 | 8.7 | 8.7 | 9.5 |
| 19 | Tumor | 44.9 | **17.1** | 23.4 | 23.4 | 36.6 | 38.1 | 44.9 |
| 20 | Chess | **30.8** | 31.5 | 31.5 | 31.5 | 31.7 | 31.7 | **30.8** |
| 21 | Letter | **1169.1** | 1200.8 | 1200.8 | 1200.9 | 1171.1 | 1171.5 | **1169.1** |
| 22 | Segment | **41.6** | 42.0 | 42.0 | 42.0 | 41.7 | 42.4 | **41.6** |
| 23 | Sick | 27.8 | **20.4** | 26.6 | 25.4 | 24.0 | 24.0 | 26.7 |
| 24 | Sonar | 14.4 | 14.5 | 14.5 | 14.5 | **13.8** | **13.8** | 14.4 |
| 25 | Splice | 175.8 | **161.4** | **161.4** | 161.9 | **161.4** | **161.4** | **161.4** |
| 26 | Anneal | 37.2 | 45.5 | 45.5 | **32.8** | 47.5 | 41.5 | 46.7 |
| 27 | Autos | 45.5 | **41.0** | 42.5 | 42.5 | 49.4 | 49.0 | 45.5 |
| 28 | Colic | 5.8 | **4.7** | 5.3 | 9.2 | 5.6 | 5.6 | 5.0 |
| 29 | Hypothyroid | 14.6 | **6.5** | 14.7 | 14.8 | 14.8 | 14.8 | 14.6 |
| 30 | Ionosphere | **14.0** | 14.3 | 14.3 | 14.3 | 14.5 | 14.7 | **14.0** |
| 31 | Labor | 4.0 | 2.7 | 3.9 | 4.8 | 3.6 | 3.6 | 4.3 |
| 32 | Lymph | 17.5 | 13.2 | 13.2 | 14.2 | **10.6** | 10.7 | 12.4 |
| 33 | Vehicle | **66.4** | 82.6 | 82.6 | 82.6 | 89.9 | 91.3 | **66.4** |
| 34 | Vote | 5.8 | **5.1** | 5.8 | 5.8 | 5.9 | 5.9 | 5.8 |
| 35 | Vowel | **130.0** | 187.1 | 187.1 | 186.5 | 186.8 | 183.3 | 186.1 |

## Appendix 2: Detailed results of analysis based on unequal misclassification costs

In Table A2 below, the ratio “1:5” implies that the misclassification cost of the first target event is 1/5 of the cost of the second target event, while the ratio “5:1” implies that the misclassification cost of the first target event is 5 times the cost of the second target event. For each dataset & ratio, the splitting method that has the best cost has the relevant value displayed in bold.

**Table A2** Performance based on unequal misclassification costs

| Dataset | Ratio of misclassification costs of target events | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Vote** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 4.9 | **4.5** | 3.4 | 2.5 | 1.7 | 2.0 | **3.0** | **3.3** | 3.7 |
| EffCAMI_1 | 4.9 | **4.5** | 3.5 | 2.5 | 1.5 | 1.8 | **3.0** | **3.3** | 3.7 |
| IG | **4.4** | 4.9 | 3.6 | 2.7 | 1.8 | 2.3 | 3.4 | 3.4 | 3.7 |
| CAMI | 4.9 | **4.5** | 3.4 | 2.5 | 1.7 | 2.0 | **3.0** | **3.3** | 3.7 |
| GR | 5.1 | **4.5** | **3.1** | **2.0** | 1.4 | **1.7** | **3.0** | **3.3** | 3.7 |
| EffCAMI_0 | 4.9 | **4.5** | 3.5 | 2.5 | 1.5 | 1.8 | **3.0** | **3.3** | 3.7 |
| adjGR | 5.1 | **4.5** | **3.1** | **2.0** | **1.4** | **1.7** | **3.0** | **3.3** | 3.7 |
| ** *Worst–Best* ** | *0.7* | *0.4* | *0.5* | *0.7* | *0.4* | *0.6* | *0.4* | *0.1* | *0.0* |
| **Breast Cancer** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 11.3 | 10.8 | 11.4 | 13.1 | 8.7 | **14.9** | 18.7 | **20.4** | **21.1** |
| EffCAMI_1 | 11.3 | 10.8 | 11.4 | 13.1 | **8.4** | 15.1 | 18.6 | 20.5 | 21.2 |
| IG | 11.3 | 10.8 | 11.4 | 13.1 | 8.7 | **14.9** | 18.7 | **20.4** | **21.1** |
| CAMI | 11.3 | 10.8 | 11.4 | 13.1 | 8.7 | **14.9** | 18.7 | **20.4** | 21.1 |
| GR | **11.1** | 10.8 | 11.4 | **12.8** | 8.9 | **14.9** | **18.5** | 22.4 | 22.7 |
| EffCAMI_0 | 11.3 | 10.8 | 11.4 | 13.1 | **8.4** | 15.1 | 18.6 | 20.5 | 21.2 |
| adjGR | **11.1** | 10.8 | 11.4 | **12.8** | 8.9 | **14.9** | **18.5** | 22.4 | 22.7 |
| ** *Worst–Best* ** | *0.2* | *0.0* | *0.0* | *0.3* | *0.5* | *0.2* | *0.2* | *2.0* | *1.6* |
| **Chess** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 7.7 | 7.1 | 5.5 | 4.5 | 1.9 | 4.1 | 5.5 | 7.0 | 8.5 |
| EffCAMI_1 | 7.7 | 7.1 | 5.5 | 4.5 | **1.7** | 3.9 | 4.9 | 6.1 | 8.1 |
| CE | 7.7 | 7.1 | 5.5 | 4.5 | 1.9 | 4.1 | 5.5 | 7.0 | 8.5 |
| CAMI | 7.7 | 7.1 | 5.5 | 4.5 | 1.9 | 4.1 | 5.5 | 7.0 | 8.5 |
| IG | **7.5** | **6.3** | 4.9 | **3.7** | **1.7** | **3.8** | **4.8** | **5.8** | **8.0** |
| EffCAMI_0 | 7.7 | 7.1 | 5.5 | 4.5 | **1.7** | 3.9 | 4.9 | 6.1 | 8.1 |
| adjGR | **7.5** | **6.3** | **4.7** | **3.7** | **1.7** | **3.8** | **4.8** | **5.8** | **8.0** |
| ** *Worst–Best* ** | *0.2* | *0.8* | *0.6* | *0.8* | *0.2* | *0.3* | *0.7* | *1.2* | *0.5* |
| **Colic** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 10.8 | 9.3 | 8.1 | **6.4** | 4.7 | 8.6 | **11.2** | **15.4** | **17.0** |
| EffCAMI_1 | 13.4 | 10.1 | 8.4 | 6.6 | **5.2** | 10.3 | 15.1 | 16.4 | 17.4 |
| IG | 13.6 | 13.6 | 13.6 | 13.6 | 8.8 | **9.6** | 11.9 | 19.2 | 22.8 |
| CAMI | **12.4** | 10.8 | 8.7 | 6.9 | 5.7 | 11.0 | 15.5 | 15.9 | 17.5 |
| GR | 11.1 | **9.1** | **7.8** | 6.5 | **5.2** | 10.7 | 15.8 | 16.9 | 17.6 |
| EffCAMI_0 | 13.4 | 10.1 | 8.4 | 6.6 | **5.2** | 10.3 | 15.1 | 16.4 | 17.4 |
| adjGR | 12.4 | 10.5 | 8.7 | 7.1 | 5.8 | 11.1 | 14.7 | 15.8 | 17.3 |
| ** *Worst–Best* ** | *2.8* | *4.5* | *5.8* | *7.2* | *4.1* | *2.4* | *4.6* | *3.8* | *5.8* |
| **Credit** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 20.0 | 19.1 | 17.0 | 14.6 | 10.4 | 15.3 | 19.2 | 19.8 | 24.1 |
| EffCAMI_1 | 18.0 | **15.9** | 14.1 | 13.4 | 11.0 | 15.3 | 19.6 | 21.5 | 22.6 |
| IG | 20.4 | 17.2 | 14.2 | **12.3** | 10.6 | 16.0 | 20.5 | 21.9 | 22.7 |
| CAMI | **16.9** | **15.9** | 14.1 | 13.2 | 11.2 | 15.3 | 19.1 | 21.6 | 22.2 |
| GR | 18.8 | 17.0 | 14.4 | 15.5 | **10.1** | **12.9** | **16.2** | **17.9** | **19.6** |
| EffCAMI_0 | 18.0 | **15.9** | 14.1 | 13.4 | 11.0 | 15.3 | 19.6 | 21.5 | 22.6 |
| adjGR | 17.6 | 16.1 | 14.9 | 13.2 | 11.0 | 14.1 | 17.4 | 18.4 | 20.1 |
| ** *Worst–Best* ** | *3.5* | *3.2* | *2.9* | *3.2* | *1.1* | *3.1* | *4.3* | *4.0* | *4.5* |
| **Heart** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 11.4 | 11.5 | 10.3 | 8.0 | 6.9 | 10.9 | 11.6 | 14.5 | 14.4 |
| EffCAMI_1 | 11.0 | 11.3 | 10.7 | 7.9 | 6.3 | 10.3 | **11.2** | 14.3 | 15.7 |
| IG | 11.4 | 11.5 | 10.3 | 8.0 | 6.9 | 10.9 | 11.6 | 14.5 | 14.4 |

*(Continued on next page.)*

**Table A2** *(Continued.)*

| Dataset | Ratio of misclassification costs of target events | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| CAMI | 11.4 | 11.5 | 10.3 | 8.0 | 6.9 | 10.9 | 11.6 | 14.5 | 14.4 |
| GR | **9.9** | **9.6** | **8.5** | **6.8** | **5.8** | **9.7** | **11.2** | **13.6** | **13.1** |
| EffCAMI_0 | 11.0 | 11.3 | 10.7 | 7.9 | 6.3 | 10.3 | **11.2** | 14.3 | 15.7 |
| adjGR | **9.9** | **9.6** | **8.5** | **6.8** | **5.8** | **9.7** | **11.2** | **13.6** | **13.1** |
| ** *Worst–Best* ** | *1.5* | *1.9* | *2.2* | *1.2* | *1.1* | *1.2* | *0.4* | *0.9* | *2.6* |
| **Hepatitis** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 8.9 | 7.6 | **6.2** | **4.4** | **2.6** | 3.2 | 3.2 | 3.0 | 3.2 |
| EffCAMI_1 | **6.6** | 7.5 | **6.2** | 5.3 | 2.9 | 3.3 | **2.6** | **2.5** | 3.2 |
| IG | 8.7 | 7.8 | 7.0 | 5.5 | 3.2 | 3.2 | 3.2 | 3.2 | 3.2 |
| CAMI | 8.9 | 7.6 | 6.2 | **4.4** | **2.6** | 3.2 | 3.2 | 3.0 | 3.2 |
| GR | 7.4 | **7.4** | **7.0** | 4.7 | 3.2 | **2.9** | **2.6** | 2.7 | 3.2 |
| EffCAMI_0 | **6.6** | 7.5 | **6.2** | 5.3 | 2.9 | 3.3 | **2.6** | **2.5** | 3.2 |
| adjGR | 7.4 | 7.4 | 7.0 | 4.7 | 3.2 | **2.9** | **2.6** | 2.7 | 3.2 |
| ** *Worst–Best* ** | *2.3* | *0.4* | *0.8* | *1.1* | *0.6* | *0.4* | *0.6* | *0.7* | *0.0* |
| **Pima** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 31.0 | 29.8 | 30.0 | 28.2 | 20.4 | 30.0 | 36.4 | 42.4 | 46.1 |
| EffCAMI_1 | 31.2 | 30.8 | 29.4 | 28.0 | 20.8 | 28.8 | 35.3 | 37.9 | 46.1 |
| IG | 31.0 | 29.8 | 30.0 | 28.2 | 20.4 | 30.0 | 36.4 | 42.4 | 46.1 |
| CAMI | 31.0 | 29.8 | 30.0 | 28.2 | 20.4 | 30.0 | 36.4 | 42.4 | 46.1 |
| GR | **27.5** | **28.6** | **25.3** | **22.2** | **17.9** | **28.5** | **32.3** | **37.1** | **41.2** |
| EffCAMI_0 | 31.2 | 30.8 | 29.4 | 28.0 | 20.8 | 29.2 | 35.3 | 37.9 | 46.1 |
| adjGR | **27.5** | **28.6** | **25.3** | **22.2** | **17.9** | **28.5** | **32.3** | **37.1** | **41.2** |
| ** *Worst–Best* ** | *3.7* | *2.2* | *4.7* | *6.0* | *2.9* | *1.5* | *4.1* | *5.3* | *4.9* |
| **Ionosphere** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 10.6 | 9.9 | 9.4 | 6.7 | **3.5** | 6.0 | **6.4** | **6.2** | **6.9** |
| EffCAMI_1 | 10.6 | **8.9** | 9.1 | 6.8 | **3.5** | **5.4** | **6.4** | **6.2** | **6.9** |
| IG | 10.6 | 9.9 | 9.4 | 6.7 | **3.5** | 6.0 | **6.4** | **6.2** | **6.9** |
| CAMI | 10.6 | 9.9 | 9.4 | 6.7 | **3.5** | 6.0 | **6.4** | **6.2** | **6.9** |
| GR | **9.5** | 10.0 | **8.9** | **6.0** | **4.0** | **6.4** | **7.1** | **6.8** | **7.1** |
| EffCAMI_0 | 10.6 | **8.9** | 9.1 | 6.8 | **3.5** | **5.4** | **6.4** | **6.2** | **6.9** |
| adjGR | **9.5** | 10.0 | **8.9** | **6.0** | **4.0** | **6.4** | **7.1** | **6.8** | **7.1** |
| ** *Worst–Best* ** | *1.1* | *1.1* | *0.5* | *0.8* | *0.5* | *1.0* | *0.7* | *0.6* | *0.2* |
| **Labor** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 3.5 | 3.6 | 2.4 | **1.9** | **1.0** | 1.5 | 1.8 | 2.3 | **3.3** |
| EffCAMI_1 | 3.2 | 2.7 | 2.3 | 2.3 | 1.2 | 1.7 | **1.6** | 2.1 | **3.3** |
| IG | 3.4 | 3.2 | 3.0 | 2.6 | 1.6 | 2.0 | 2.0 | **2.0** | **2.0** |
| CAMI | 3.4 | 3.6 | 2.7 | 2.2 | 1.3 | 1.6 | 1.9 | 2.3 | **3.3** |
| GR | 3.3 | 2.8 | **2.2** | **1.9** | 1.2 | **1.3** | 1.8 | 2.4 | **3.3** |
| EffCAMI_0 | 3.2 | 2.7 | 2.3 | 2.3 | 1.2 | 1.7 | **1.6** | 2.1 | **3.3** |
| adjGR | **2.6** | **2.5** | 2.6 | 2.5 | 1.6 | 1.7 | 2.1 | 3.0 | **3.3** |
| ** *Worst–Best* ** | *0.3* | *0.9* | *0.8* | *0.7* | *0.6* | *0.7* | *0.4* | *0.4* | *1.3* |
| **Sick** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | 7.8 | 6.9 | **6.1** | 6.1 | 3.7 | 6.1 | 7.9 | 9.4 | 11.3 |
| EffCAMI_1 | 8.1 | 7.6 | 7.6 | **5.4** | 4.3 | 5.0 | 6.4 | 8.8 | 10.5 |
| IG | 7.0 | 6.8 | 6.8 | 5.9 | 5.8 | 10.4 | 11.9 | 13.9 | 16.7 |
| CAMI | **6.9** | 6.7 | 7.6 | 5.6 | **3.4** | 6.1 | 7.4 | 8.8 | 11.6 |
| GR | **6.9** | **6.6** | 6.6 | 6.0 | 4.9 | 5.1 | **5.0** | 8.7 | 10.3 |
| EffCAMI_0 | 8.1 | 7.6 | 7.6 | **5.4** | 4.3 | 5.0 | 6.4 | 8.8 | 10.5 |
| adjGR | **6.9** | **6.6** | 6.6 | 5.6 | 4.8 | **4.8** | **5.0** | **7.9** | **9.7** |
| ** *Worst–Best* ** | *1.2* | *1.0* | *1.5* | *0.7* | *2.4* | *5.4* | *6.9* | *5.2* | *6.4* |

*(Continued on next page.)*

**Table A2** *(Continued.)*

| Dataset | Ratio of misclassification costs of target events | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sonar** | **1/5** | **1/4** | **1/4** | **1/2** | **1/1** | **2/1** | **3/1** | **4/1** | **5/1** |
| CAIR | **11.7** | **10.1** | **7.3** | 7.3 | 5.7 | 9.8 | 12.7 | 12.2 | 14.3 |
| EffCAMI_1 | 13.9 | 13.8 | 9.3 | 6.8 | 5.6 | **7.3** | **8.8** | **9.0** | 12.4 |
| IG | **11.7** | **10.1** | **7.3** | 7.3 | 5.7 | 9.8 | 12.7 | 12.2 | 14.3 |
| CAMI | **11.7** | **10.1** | **7.3** | 7.3 | 5.7 | 9.8 | 12.7 | 12.2 | 14.3 |
| GR | 12.1 | 11.0 | 8.5 | **6.3** | 5.7 | 8.7 | 11.6 | 11.1 | **9.3** |
| EffCAMI_0 | 13.9 | 13.8 | 9.3 | 6.8 | 5.6 | **7.3** | **8.8** | **9.0** | 12.4 |
| adjGR | 12.1 | 11.0 | 8.5 | **6.3** | 5.7 | 8.7 | 11.6 | 11.1 | **9.3** |
| ** *Worst–Best* ** | *2.2* | *3.7* | *2.0* | *1.0* | *0.1* | *2.5* | *3.9* | *3.2* | *5.0* |

**Acknowledgments** We thank the anonymous referees for their valuable comments that enabled us to improve the quality of this paper. This research was supported in part by a grant from the 2004 Summer Research Program of the School of Business of Virginia Commonwealth University.

## References

Bradley P, Fayyad Usama M, Mangasarian OL. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing* 1999;11(3):217–238.

Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Wadsworth, California, USA, 1984.

Bryson K-M. On two families of entropy-based splitting methods. Working Paper, Department of Information Systems. Virginia Commonwealth University, USA, 2000.

Cheeseman P, Stutz J, Bayesian Classification (AutoClass): Theory and results. In: Gregory Piatetsky-Shapiro Usama Fayyad, Padhraic Smyth, ed, *Advances in Knowledge Discovery and Data Mining*, Menlo Park, AAAI Press, MIT Press, 1996; 153–180.

Ching J, Wong A, Chan K. Class-dependent discretization for inductive learning from continuous and mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1995;17(7):631–641.

Gersten W, Wirth D, Arndt D. Predictive modeling in automative direct marketing: Tools, experiences and open issues. In: *Proceedings of 2000 International Conference on Knowledge Discovery and Data Mining (KDD-2000)* Boston, MA, 2000; 398–406.

Lopez de Mantaras R. A Distance-based attribute selection measure for decision tree induction. *Machine Learning* 6: 1991; 81–92.

Martin J. An exact probability metric for decision tree splitting and stopping. *Machine Learning* 1997;28:257–291.

Murphy P, Aha DW. *UCI Repository of Machine Learning Databases*. University of California, Department of Information and Computer Science,1994.

Piatetsky-Shapiro G. The data-mining industry coming of age. *IEEE Intelligent Systems* 1999;14(6):32–34.

Quinlan J. Induction of decision trees. *Machine Learning* 1986;1:81–106.

Quinlan J. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

Safavian S, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 1991;21(3):660–674.

Shih Y-S. Families of splitting criteria for classification trees. *Statistics and Computing* 9:4, 1999; 309–315.

Taylor P, Silverman B. Block diagrams and splitting criteria for classification trees. *Statistics and Computing* 3(4), 1993; 147–161.

Wu X, Urpani D. Induction by attribute elimination. *IEEE Transactions on Knowledge and Data Engineering* 1999; 11(5):805–812.

**Kweku-Mauta Osei-Bryson** is Professor of Information Systems at Virginia Commonwealth University, where he also served as the Coordinator of the Ph.D. program in Information Systems during 2001–2003. Previously he was Professor of Information Systems and Decision Analysis in the School of Business at Howard University, Washington, DC, U.S.A. He has also worked as an Information Systems practitioner in both industry and government. He holds a Ph.D. in Applied Mathematics (Management Science & Information Systems) from the University of Maryland at College Park, a M.S. in Systems Engineering from Howard University, and a B.Sc. in Natural Sciences from the University of the West Indies at Mona. He currently does research in various areas including: Data Mining, Expert Systems, Decision Support Systems, Group Support Systems, Information Systems Outsourcing, Multi-Criteria Decision Analysis. His papers have been published in various journals including: Information & Management, Information Systems Journal, Information Systems Frontiers, Business Process Management Journal, International Journal of Intelligent Systems, IEEE Transactions on Knowledge & Data Engineering, Data & Knowledge Engineering, Information & Software Technology, Decision Support Systems, Information Processing and Management, Computers & Operations Research, European Journal of Operational Research, Journal of the Operational Research Society, Journal of the Association for Information Systems, Journal of Multi-Criteria Decision Analysis, Applications of Management Science. Currently he serves an Associate Editor of the INFORMS Journal on Computing, and is a member of the Editorial Board of the Computers & Operations Research journal.

**Kendall E. Giles** received the BS degree in Electrical Engineering from Virginia Tech in 1991, the MS degree in Electrical Engineering from Purdue University in 1993, the MS degree in Information Systems from Virginia Commonwealth University in 2002, and the MS degree in Computer Science from Johns Hopkins University in 2004. Currently he is a PhD student (ABD) in Computer Science at Johns Hopkins, and is a Research Assistant in the Applied Mathematics and Statistics department. He has over 15 years of work experience in industry, government, and academic institutions. His research interests can be partially summarized by the following keywords: network security, mathematical modeling, pattern classification, and high dimensional data analysis.