

## Mining Data to Find Subsets of High Activity

Dhammika Amaratunga<sup>1</sup> and Javier Cabrera<sup>2</sup>

<sup>1</sup>The R.W. Johnson Pharmaceutical Research Institute, Raritan, NJ 08869  
(damaratu@prius.jnj.com)

<sup>2</sup>National University of Singapore and Rutgers University  
(cabrera@stat.rutgers.edu)

### ABSTRACT

Many data mining problems are concerned with trying to identify the characteristics of a subset of cases that responds substantially differently from the rest of the cases. For example, when studying the relationship between a response variable  $Y$  and a set of predictor variables, it is often of interest to determine what ranges of values of the descriptor variables are associated with a high likelihood of  $Y=1$  (if  $Y$  is a Bernoulli variable) or with high values of  $Y$  (if  $Y$  is a continuous variable). We describe a criterion ( $H$ ) and a recursive partitioning method (ARF) that directly addresses this question. A computational algorithm that makes ARF feasible for use even with very large data mining datasets is presented. The basic version of ARF can be generalized to the case of multiple response variables,  $Y_1, \dots, Y_t$  and other settings. Mining a structure activity database (and some other datasets) illustrates the effectiveness of the ARF procedure. We conclude by proposing a basic paradigm for data mining.

KEY WORDS: ARF, data mining, recursive partitioning, classification tree.

### 1. INTRODUCTION

Data mining is the computer-intensive activity of exploring large datasets in the hope of discovering, within a subset of the data, some relationship or pattern or hypothesis worthy of further study. These data are typically very messy, *i.e.*, noisy, have missing values, have extreme outliers, and are perhaps far from Gaussian, even multimodal. More critically, as the data are often observational or opportunistic, they are likely to be heterogeneous, so relationships between variables in the dataset may not necessarily be the same across all the data, and nuggets of interest may be buried within some subsets of variables and cases. All this makes data mining a challenge (Fayyad *et al* (1996) is an extensive introduction to data mining; Friedman (1997) discusses the interplay between statistics and data mining).

Like its close relative, exploratory data analysis (Tukey, 1962, 1977), data mining is best guided by objectives, however vague. A principal objective of many data mining problems is to uncover characteristics of subsets of cases that respond substantially differently from the rest of the cases. Consider these case studies.

*Case Study 1:* Structure activity databases (SADB) in the pharmaceutical industry are datasets prepared with the objective of studying the relationship between the biological activities of a series of compounds and their chemical properties. A particular anxiolytic SADB (internal dataset) that we studied had the form  $\{(Y_i, x_i)\}$ , where  $i=1, \dots, N$  indexes the compounds in the database. The  $t$ -vector,  $Y_i=(Y_{i1}, \dots, Y_{it})'$ , was a set of  $t$  binary variables indicating the results of running the  $i$ th compound through a series of  $t$  *in vivo* biological assays,  $Y_{ji}=1$  or 0 depending on whether or not the  $i$ th compound was active in

the  $j$ th assay. The  $r$ -vector  $x_j=(x_{j1},\dots,x_{jr})'$  was a collection of  $r$  predictor variables, consisting of measurements of physicochemical properties, including compound class, molecular weight, molecular volume, logP and the number of rotatable bonds, as well as *in vitro* measurements of activity at the cellular level such as IC50.

The primary goal in analyzing this data was to identify ranges of values of  $x=(x_1,\dots,x_r)$  associated with higher likelihoods of *in vivo* activity. That is, for any assay,  $j$ , we wish to find subsets of the form  $S=\prod I_k$ , where  $I_k$  is an interval of values of variable  $x_k$ , that have a relatively high value of  $P[Y_j=1|S]$ . We refer to such subsets as "high activity regions" (HARs). A key consideration here is that subsets of the form  $S=\prod I_k$  (which may refer to, *e.g.*, compounds with a certain range of molecular weights and IC50s) are simpler to communicate to nonstatisticians and are more easily understood and interpreted by them, as opposed to subsets of a more general nature.

Different assays will, of course, tend to have different HARs. Thus, of equal interest was to identify "high selectivity regions" (HSRs), ranges of values of  $x$  that simultaneously exhibit high activity in *in vivo* efficacy assays and low activity in *in vivo* toxicity assays. Here we will consider one efficacy assay, A1, and one toxicity assay, A2.

Initial exploration of the SADB data revealed features that would complicate any analysis of it: (a) A nonnegligible fraction of the *in vivo* activity data (the  $\{Y_{ji}\}$ ) and the *in vitro* activity data was missing. (b) There were a number of extreme outliers among the  $\{x_{ki}\}$ . (c) There were many strong dependencies among the  $\{x_k\}$ , most notably, among the various measurements related to molecular size, such as weight, volume, and the number of rotatable bonds. (d) Scatterplots (see Fig 1) indicated that the relationships between  $P[Y_{ij}=1]$  and  $x_{ki}$  (or more precisely  $P[Y_{ij}=1 | c-\epsilon < x_{ki} < c+\epsilon]$  vs  $c$  for small values of  $\epsilon$ ) were nonmonotone for most  $j$  and  $k$ . (e) There were many response variables; while they could be studied individually to identify HARs, they have to be studied in combination to identify HSRs.

**Table 1:** The "toy" dataset - a subset of the SADB.

MOLWT	IC50	Y1	Y2
256	10000.00	0	0
266	NA	0	0
294	2080.00	0	NA
319	10000.00	0	0
324	2.79	0	0
335	9.94	1	0
335	26.30	1	0
340	122.00	NA	0
352	246.60	1	0
353	8.40	1	0
354	0.96	1	1
359	0.14	1	NA
360	NA	1	1
376	2.40	1	1
378	11.01	1	1
387	1.79	0	1
401	16.50	0	1
417	10.65	NA	0
421	101.59	0	0
470	10000.00	0	0

MOLWT = Molecular weight; IC50 = IC50; Y1 = Activity observed in assay A1; Y2 = Activity observed in assay A2; NA = Missing value.

It is instructive to examine the toy dataset based on the SADB shown in Table 1. Despite outliers and missing values in the data, it is clear that assay A1 is generally active for compounds with molecular weight (MOLWT) in the interval [335, 378] and IC50s 26.30 and below, assay A2 is generally active for compounds with MOLWT in [354, 401] and IC50s 16.50 and below, and both assays are simultaneously active for compounds with MOLWT in [354, 378] and IC50s 16.50 and below. While such patterns are easy to spot in a tiny dataset like this, it is considerably more challenging to do so in one many times larger.

*Case Study 2:* A company that produces and sells medical instruments (internal dataset) prepared a database on the sales of orthopedic material to hospitals. The data was of the form  $\{(Y_i, x_i)\}$ , where  $i=1, \dots, N$  indexes hospitals;  $Y_i$  was the log of the sales of rehabilitation equipment for 1996 and  $x_i$  was a collection of 15 predictor variables, consisting of information for each hospital, such as the number of beds, administrative costs, inpatient revenue, the number of hip and knee and femur operations for the past two years, whether or not it is a teaching hospital, whether or not it has a trauma unit, city and state. The objective was to identify the characteristics of those hospitals with high sales volumes.

*Case Study 3:* An epidemiological database of several women of Pima Indian heritage, collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, studied the relationship between the incidence of diabetes among them and several predictor variables, such as age, plasma concentration level, serum insulin level, diastolic blood pressure, body mass index (Blake and Merz (1998) has the dataset). The objective was to identify the characteristics of the subjects associated with a high incidence rate of diabetes.

*Case Study 4:* A clinical trial (internal dataset) showed that a certain drug was efficacious, compared to placebo, for treatment of anemia. A "subgroup analysis" was requested to study whether the patients who responded positively (*i.e.*, did not need a blood transfusion) in the treatment group differed from those who responded positively in the placebo group in terms of their baseline levels of hematology parameters, particularly hemoglobin.

The common factor in all these case studies is that the question being asked is "what are the characteristics of the subset of cases that respond positively (or have high response values)?" In this paper, we describe an approach that directly addresses this question. We present novel methodology that builds on existing methods for classification and regression trees such as CART (Breiman *et al* (1984)), CHAID (Hartigan (1975)), FIRM (Hawkins and Kass (1982)), and, more recently, Lee and Buja (1999). Over the next several sections, we will describe our method and discuss its performance with simulations and examples.

## 2. THE $H$ CRITERION

The basis of our approach is the  $H$  criterion, which we now describe for data of the form  $D=\{(Y_i, x_i)\}$ , where  $i=1, \dots, N$ ,  $Y_i$  is a Bernoulli variable, which is either 0 ("failure") or 1 ("success"), and  $x_i=(x_{1i}, \dots, x_{ri})'$  is an  $r$ -vector of predictor variables. The objective is to discover HARs, ranges of values of  $x$  (*i.e.*, subsets of the form  $S=\prod_k$ , as described above) associated with high values of  $\text{Prob}(Y=1|S)$  (*i.e.*, with high success probabilities).

## 2.1 The criterion

It is natural to consider that, for the  $k$ th predictor variable,  $x_k$ , an ‘interesting’ interval,  $I_k = \{a_k \leq x_k \leq b_k\}$ , is one that has a substantially higher proportion of successes compared to  $D$ , *i.e.*, one such that  $p(I_k) = \text{Prob}(Y=1|I_k)$  is substantially larger than  $p(D) = \text{Prob}(Y=1|D)$ . In order to compare  $p(I_k)$  across subsets,  $I_k$ , of different sizes on an equal footing, we need a statistic that is not much dependent on  $n(I_k)$ , the number of observations in  $I_k$ . Such a statistic is

$$z(I_k;D) = (p(I_k) - p(D)) / \sigma_p, \quad \text{where } \sigma_p^2 = p(D)(1-p(D))/n(I_k),$$

as  $z(I_k;D)$  is approximately  $N(0,1)$ , irrespective of sample size, except for very small samples, for a random binary series of length  $N$  with success probability  $p(D)$ . The larger the value of  $z(I_k;D)$ , the more interesting is  $I_k$ .

We could use  $z(I_k;D)$  as the criterion for identifying potential HARs and, in fact, it works quite well in this regard (see Section 5). Sometimes, however, a large interval with a large proportion of successes, but with a few failures in the middle, gets split at these failures due to one of the subintervals having a slightly higher  $z$ . In such cases, it is desirable to capture the large (undivided) interval with the intention of splitting it later if necessary. This can be attained by adding, to  $z(I_k;D)$ , a penalty term,  $\psi$ , that is small for small subsets. A natural choice for  $\psi$  is  $\lambda \log(h)$ , where  $h = n(I_k)$  and  $\lambda$  is a prespecified constant; dividing this by  $\log(N)$  reduces the dependence of  $\lambda$  on  $N$ , making  $\lambda$  more interpretable when using the criterion recursively. Thus, our criterion for judging how interesting a subset is is the *H criterion*:

$$H(I_k;D) = z(I_k;D) + \lambda \log(h) / \log(N).$$

As with  $z(I_k;D)$ , the larger the value of  $H(I_k;D)$ , the more interesting the subset.

We have found it reasonable to set  $\lambda$  to 5 on most occasions (see Section 5).

*Example:* Some results of applying the *H* criterion with  $\lambda=5$  to find HARs for the SADB are displayed in Figure 1. Assay A1 results (30% overall success rate) are plotted against each of three predictor variables and a lowess smooth overlaid on each plot. The vertical lines in the plots are the intervals that have the largest value of  $H$  for that variable. For comparison, the column on the right gives the standard SPlus classification trees after pruning. IC50 provides a typical example of how segmentation methods behave when the relationship is monotone; the splits are similar and reflect the monotone structure;  $H$  picks out a 26% left interval with a success rate of 86%. With DM (total dipole moment), which has a more complicated relationship,  $H$  picks out an internal 32% interval with a 38% success rate, unlike the standard tree which peels the data from right to left without ever revealing this piece. MOLWT (molecular weight) shows a quadratic relationship with medium molecules having the highest activity; a central 24% interval of high (57%) activity is immediately identified by  $H$ ; more or less the same interval is detected by the standard tree, but only after three splits.

*Remark:* Focusing on interval subsets (as opposed to the more common practice of splitting the data into two at some value in the range of  $x_k$ ) allows quadratic and other nonmonotone segments of the relationship between the activity rate  $\text{Prob}[Y=1]$  and  $x_k$  to be identified, an important consideration since many relationships exhibit this type of shape if the range of  $x$  is sufficiently broad. For example, (a) as we have already seen in Case Study 1, with structure activity data, medium sized molecules are more likely to be highly active than small or large ones, (b) in Case Study 4, patients with low hemoglobin

levels were too compromised to respond to either drug or placebo while those with high hemoglobin levels responded with or without treatment, so that differentiation between drug and placebo occurs in the mid-range of baseline hemoglobin levels.

*Remark:* Nominal categorical predictor variables (e.g., compound class in Case Study 1 and state in Case Study 2) are handled slightly differently. Rather than an interval, we seek a subset of the levels of the variable such that the criterion is optimized.

*Remark:* To avoid favoring very large subsets, we could force  $\psi$  to remain constant once it has reached a threshold by setting  $\psi = \lambda \min\{\log(h), \log(fN)\} / \log(fN)$ , where  $f$  is the specified fraction of data corresponding to the threshold subset size. Whether this offers a substantial improvement is unclear and is still under investigation.

## 2.2 Assessing the importance of a subset

The value of  $H$  can be used to assess the importance of a subset  $I_{max}$  selected by maximizing it. We regard  $I_{max}$  as important if  $H(I_{max}) > h(N, \pi, \lambda)$ , where  $h(N, \pi, \lambda)$  is the 95<sup>th</sup> percentile of the distribution of  $H$  in a random binary series of length  $N$  with success probability  $\pi = p(D)$ . We performed a simulation to determine  $h(N, \pi, \lambda)$  for  $\lambda = 5$ ,  $\pi = 0.1, 0.2, \dots, 0.9$  and  $N = 100, 500, 1000, 2000, 5000, 10000$ . For each case we did 2000 simulations and the resulting 95th percentiles were obtained. A very good approximation for  $h(N, \pi, \lambda)$  was found to be given by

$$h(N, \pi, \lambda) = -0.5 + 0.7(\pi + \log(N)) - 0.3\pi \log(N).$$

In addition, the longest sequence of successes in a random binary series is known to be  $O(\log N)$  (see Gordon, Schilling and Waterman (1986), Schilling (1990)). We found through simulation that a very good approximation to the 95<sup>th</sup> percentile of the longest run of successes in a binary series with success probability  $\pi$  is  $-1.33 \log(N) / \log(\pi)$ . Therefore we ignore subsets smaller than this.

*Example:* According to the  $h(N, \pi, \lambda)$  critical value, for the SADB data with assay A1 results as response, the interval  $IC_{50} < 7.11$  selected for  $IC_{50}$  is significant. So is the interval  $[448, 683]$  for MOLWT. The interval  $[5.86, 7.86]$  for DM is nonsignificant.

## 3. RECURSIVE PARTITIONING WITH THE $H$ CRITERION

Since the  $H$  criterion is specifically designed to find HARs, we use the acronym ARF ("Activity Region Finder") to refer to the following recursive partitioning procedures that use it.

### 3.1 Procedure ARF1

*Procedure ARF1:* To mine a dataset  $D_0$  of size  $n(D_0)$ :

STEP 1: Find, for each  $x_k$ , the interval,  $I_k$ , that maximizes  $H(I_k; D_0)$ . Search among these intervals for the one that has the largest value of  $H(I_k; D_0)$ ; say this is the one associated with variable  $x_{k1}$  and interval  $I_{k1} = \{a_{k1} \leq x_{k1} \leq b_{k1}\}$ .

STEP 2 Repeat the process with the subset  $D_1 = \{(Y_i, x_i) \in D_0 : x_{k1} \in I_{k1}\}$ . Let the variable and interval selected at the second step be  $x_{k2}$  and  $I_{k2}$ .

...

STEP  $s$ : Repeat the process with the subset  $D_s = \{(Y_i, x_i) \in D_{s-1} : x_{k,s-1} \in I_{k,s-1}\}$ . Let the variable and interval selected at this step be  $x_{ks}$  and  $I_{ks}$ .

STOPPING RULE: The process terminates (at step  $s'$  say) when

$$n(I_k) < -1.33 \log(N(D_0)) / \log(p(D_0))$$

(see Section 2.2). This last subset is not used since it is too small.

Once the process terminates, we use the distribution of  $H$  as a guideline to judge the "significance" of the finding at each step  $s$  (see Section 2.2). Let  $I_{s^*}$  be the first interval for which  $H(I_{ks}; D_{s-1}) > h(n(D_{s-1}), p(D_{s-1}), \lambda)$  when we look at the results from step  $s'-1$  to step 1. The sequence of subsets defined by  $I_{k1}, I_{k2}, I_{k3}, \dots, I_{ks^*}$  is a HAR. The nature of the  $H$  criterion is such that generally  $s^*$  tends to be small.

*Example:* Running ARF1 on the SADB with assay A1 responses as the response variable gives the result shown on the top of Table 2, which indicates  $IC_{50} < 7.11$  and  $DM \in [6.06, 27.64]$  as the HAR, with an activity rate of over 95%, compared to the background rate of 30%.

Table 2: ARF1 Results

Resp	Step	Var	Min	Lower	Upper	Max	%Success	%Interval	H	N	Sf
A1	1	IC50	0.07	0.07	7.11	61600.00	86.41	26.14	16.05	184	S
	2	DM	1.55	6.06	27.64	35.82	95.50	59.78	2.50	111	S
	3	VOL	231.50	260.60	315.50	427.40	100.00	47.75	0.92	54	NS
A2	1	IC50	0.07	12.80	61600.00	61600.00	96.46	66.62	7.71	452	S
	2	IC50	12.80	805.00	61600.00	61600.00	99.07	47.12	1.69	214	S
	3	LOGP	-0.57	1.72	5.03	5.03	100.00	71.96	0.37	155	NS
AS	1	IC50	0.07	1.52	23.70	61600.00	36.41	27.56	7.96	184	S
	2	ENERGY	13.74	50.76	151.25	151.25	66.67	12.50	2.58	24	S
DIAB	1	PLASMA	0.00	144.00	199.00	199.00	71.591	22.7865	13.913	176	S
	2	PLASMA	144.00	167.00	199.00	199.00	86.076	44.3182	6.799	79	S
	3	BODY	22.90	29.70	45.40	59.40	90.909	82.2785	5.544	66	NS

Resp is the response variable; Var is the variable  $x_k$  selected at step  $i$ ; Min and Max are the minimum and maximum values of  $x_k$  at step  $i$ ; Lower and Upper defines the interval  $I_k$  selected; %Success=percentage of Successes in  $I_k$ , %Interval=percentage of observations in  $I_k$ ; H=value of the  $H$  criterion; N=number of observations; Sf=S if  $H > h(n, \pi, \lambda)$  and NS otherwise.

The finding that compounds that are highly active *in vitro* (i.e., with low  $IC_{50}$ ) are also likely to be efficacious *in vivo* is hardly surprising. But high *in vitro* activity can also presage toxicity and this happens with this data, as can be seen by running ARF1 with toxicity assay A2 responses (see Table 2). To find a high selectivity region (HSR) for this data, let AS denote selectivity, indicating a success in assay A1 and a failure in assay A2, and run ARF1 with  $A_s$  as the response variable (see Table 2). This now indicates an internal interval [1.52, 23.70] for  $IC_{50}$  and [50.76, 151.25] for ENERGY as the HSR, with an activity rate of over 66%, compared to the background rate of only 15% (see Section 7 for a different analysis of this data).

*Example:* The result of running ARF1 in Case Study 3 is shown in Table 2.

*Remark:* The result of ARF1 can be shown as a tree (see Figure 2).

### 3.2 Procedure ARF2

*Procedure ARF2:* The other procedure is, at each step  $(s+1)$ , for  $s=1,2,\dots$ , instead of only following the subset  $\{(Y_i, x_i) \in D_{s-1} : x_{ks} \in I_{ks}\}$ , to follow all three subsets,  $I_{ks} = \{(Y_i, x_i) \in D_{s-1} : x_{ks} \in I_{ks}\}$ ,  $I_{ks}^- = \{(Y_i, x_i) \in D_{s-1} : x_{ks} < a_{ks}\}$  and  $I_{ks}^+ = \{(Y_i, x_i) \in D_{s-1} : x_{ks} > b_{ks}\}$ .

The results of these procedures can be shown as trees with splits at the nodes.

*Example:* Running ARF2 on the SADB with assay A1 responses as the response gives the result shown in Table 3 and Figure 3. Now, three other subsets, besides the one identified by ARF1, look somewhat interesting: (1) CLC:  $IC50 \leq 7.11$ ,  $DM \geq 1.55$ ,  $1.53 \leq GShift \leq 3.27$ , (2) RC:  $9.28 \leq IC50 \leq 35.00$ , (3) RCC:  $9.28 \leq IC50 \leq 35.00$ ,  $1.83 \leq \log P \leq 2.56$ .

Table 3: Results of running ARF2 with assay A1 results as response.

Node	Var	Min	Lower	Upper	Max	%Success	%Interval	Crit	N	Sf	NdTy
C	IC50	0.07	0.07	7.11	61600.00	86.41	26.14	16.05	184	S	FU
CC	DM	1.55	6.06	27.64	35.82	95.50	59.78	2.50	111	S	FU
CCC	VOL	231.50	260.60	315.50	427.40	100.00	47.75	0.92	54	NS	TE
CCL	VOL	231.50	231.50	260.60	427.40	91.67	3.25	0.00	24	.	TE
CCR	VOL	231.50	315.50	427.40	427.40	90.91	4.47	0.00	33	.	TE
CL	DM	1.55	1.55	6.06	35.82	76.92	8.80	0.00	65	.	FU
CLC	GShift	0.82	1.53	3.27	4.56	93.94	49.23	1.87	33	S	TE
CLL	GShift	0.82	0.82	1.53	4.56	62.96	3.65	0.00	27	.	TE
CLR	GShift	0.82	3.27	4.56	4.56	40.00	0.68	0.00	5	.	TE
CR	DM	1.55	27.64	35.82	35.82	37.50	1.08	0.00	8	.	TE
R	IC50	0.07	7.11	61600.00	61600.00	11.24	69.82	0.00	516	.	FU
RC	IC50	7.28	9.28	35.00	61600.00	40.74	15.50	8.00	81	S	FU
RCC	LOGP	1.26	1.83	2.56	4.74	66.67	32.10	2.29	27	S	TE
RCL	LOGP	1.26	1.26	1.83	4.74	9.09	1.49	0.00	11	.	TE
RCR	LOGP	1.26	2.56	4.74	4.74	32.56	5.82	0.00	43	.	TE
RL	IC50	7.28	7.28	9.28	61600.00	22.22	2.44	0.00	18	.	TE
RR	IC50	7.28	35.00	61600.00	61600.00	5.04	56.43	0.00	417	.	FU
RRC	IC50	36.20	65.40	126.00	61600.00	24.44	10.55	5.20	45	S	TE
RRL	IC50	36.20	36.20	65.40	61600.00	7.50	5.41	0.00	40	.	TE
RRR	IC50	36.20	126.00	61600.00	61600.00	2.11	44.93	0.00	332	NS	FU
RRRCL	IC50	132.00	167.00	596.00	61600.00	7.50	23.80	2.55	80	.	TE
RRRCL	IC50	132.00	132.00	167.00	61600.00	0.00	1.89	0.00	14	.	TE
RRRCL	IC50	132.00	596.00	61600.00	61600.00	0.42	32.21	0.00	238	.	TE

Node: The interval  $I_{ks}$  selected at each nonterminal node is denoted C.  $I_{ks}^-$  is denoted L and  $I_{ks}^+$  is denoted R. The letters in the node describe, from right to left, the subset selected (e.g., CR refers to the path: R at the first step,  $IC50 > 7.11$ , C at the second step,  $27.54 < DM < 35.82$ ). See Table 2 for descriptions of the other columns.

*Remark:* If the  $Y$  vs  $x_j$  relationship has more than one HAR, ARF2 should reveal the multiple HARs.

## 4. RECURSIVE PARTITIONING ALTERNATIVES

A nice feature of the ARF procedures is that they identify a potential HAR in very few steps of recursion. This is because the  $H$  criterion being maximized focuses entirely on the incidence rate of successes within the interval, so that the subset selected at each step will have within it a substantial (and rapidly growing) proportion of successes.

Statisticians have been growing trees at least since Sonquist and Morgan (1963). Trees have mostly been used for classification. Many methods for growing classification trees have been proposed by statisticians (e.g., CHAID (Hartigan (1975)), FIRM (Hawkins and Kass (1982)), CART (Breiman et al, 1984), SPlus TREE (Clark and Pregibon (1992))) and computer scientists (e.g., C4.5 (Quinlan (1993))). These

conventional recursive partitioning methods (CRPMs) generate partitions of the  $\{x_k\}$  with the goal of reaching a partition that in some sense explains or predicts  $Y$ .

Parallels can be drawn between CRPMs and ARF. However, their objective and that of ARF are not quite the same. CRPMs were developed in the spirit of classification and aim to produce a partition whose classes predict the response (*i.e.*, classify the cases) in some optimal sense. ARF, on the other hand, tries to find regions where the concentration of successes is highest. Consequently, CRPMs treats failures and successes equally, whereas ARF focuses only on the successes.

CRPMs can, of course, be used to find HARs, but do not perform all that well when used for this purpose. They tend to produce large complex trees when the dataset contains many variables, particularly when there are strong dependencies among the variables and when the  $P[Y=1]$  vs  $x_k$  relationships are nonmonotone (both of which happen in the SADB). As a consequence, HARs tend to be found far down the tree and such regions can be rather ambiguous and quite difficult to interpret. With datasets like the SADB having only a small percentage of successes concentrated in a small region and some others scattered around, CRPMs slowly peel away progressively smaller pieces of the data, again producing an elaborate tree that can be hard to decipher. On the other hand, ARF will ignore sparse successes and home in on regions where they are concentrated, so that the partitions that evolve out of even just one or two ARF steps already deserve a more careful look, which is very important, because such findings are much more readily interpretable.

Buja and Lee (1999)'s data mining criterion for classification, like ARF, focuses on successes, and was our inspiration for the  $H$  criterion. However, this does not penalize small subsets and uses binary splits.

*Remark:* The interval splits of ARF differ from the ternary splits of CRPMs. The  $z$  criterion is not equivalent to the criteria used in those methods because the only data that are used to calculate  $z$  are the data within the subset. Also, they do not include a penalty.

*Remark:* ARF produces trees that more closely resemble real trees than the trees grown by other recursive partitioning methods. An ARF tree has a trunk, out of which grow branches, each of which is itself like the main trunk of a new tree; this goes on and on up to the leaves. Other recursive partitioning methods produce trees whose branches cannot be considered equal.

## 5. SIMULATION RESULTS

Two questions of interest regarding the  $H$  criterion, (1) the choice of  $\lambda$ , (2) how  $H$  performs when used recursively, were studied via simulation.

### 5.1 The choice of $\lambda$

We carried out the following simulation to address the first of these questions. We ran 2000 simulations with  $\lambda$  varying from 0 to 12. Each sample,  $D$ , consisted of 1000 observations with the  $x$  values equally spaced between -1 and 1. The target interval  $I_T$  was set to be the interval of  $x$ 's between -0.5 to 0 (a 25% interval).

We calculate, from the interval  $I$  produced by  $H$ , the percentage  $Q_1$  of the true interval  $[-0.5, 0]$  that is covered by  $I$  and the percentage  $Q_2$  of  $I$  that does not cover the true interval. We say that the coverage of  $I$  is "Good" when  $Q_1 > 75\%$  and  $Q_2 < 25\%$ , "UnderCover" when  $Q_1 < 75\%$  and  $Q_2 < 25\%$ , "OverCover" when  $Q_1 > 75\%$  and  $Q_2 > 25\%$ ,

and "Fail" when  $Q_1 < 75\%$  and  $Q_2 > 25\%$ . The truly important measure is "Good". OverCover and UnderCover are reported as they do give information, but OverCover in some cases could correspond to a very large interval which is tantamount to a "Fail" and UnderCover could correspond to a very small interval that would be irrelevant. The percentage of runs that fell into the various categories are shown in Table 4.

Table 4a: Performance of the  $H$  criterion when  $P(Y=1|D-I_T)=0.3$ .

$\lambda$	$P(Y=1 I_T)=0.60$				$P(Y=1 I_T)=0.40$			
	UC	OC	F1	Gd	UC	OC	F1	Gd
12	0.7	0.7	0	98.6	1.8	77.6	1.9	18.7
10	1.2	0.3	0	98.5	3.7	66.1	3.0	27.2
8	1.5	0.1	0	98.4	7.8	52.4	4.3	35.5
6	2.2	0	0	97.8	13.1	37.1	6.7	43.1
5	2.7	0	0	97.3	18.9	29.7	7.8	43.6
4	3.3	0	0	96.7	24.8	22.9	8.4	43.9
3	4.0	0	0	96.0	31.4	15.6	10.5	42.5
1	5.7	0	0	94.3	49.5	5.6	13.9	31.0
0	6.4	0	0	93.6	58.5	2.8	14.8	23.9

Table 4b: Performance of the  $H$  criterion when  $P(Y=1|D-I_T)=0.15$ .

$\lambda$	$P(Y=1 I_T)=0.45$				$P(Y=1 I_T)=0.30$				$P(Y=1 I_T)=0.20$			
	UC	OC	F1	Gd	UC	OC	F1	Gd	UC	OC	F1	Gd
12	0.5	0.1	0	99.4	3.6	20.8	0.3	75.3	0.6	86.8	5.1	7.5
10	0.7	0	0	99.3	5.9	14.2	0.4	79.5	2.5	78.5	7.3	11.7
8	1.4	0	0	98.6	9.2	9.7	0.3	80.8	7.7	65.3	10.0	17.0
6	1.7	0	0	98.3	13.4	6.5	0.1	80.0	17.0	46.1	15.5	21.4
5	2.2	0	0	97.8	16.4	5.0	0.1	78.5	24.7	34.8	18.8	21.7
4	3.1	0	0	96.9	20.1	3.5	0.3	76.1	32.7	24.0	21.5	21.8
3	3.7	0	0	96.3	25.1	2.3	0.5	72.1	40.4	15.1	25.0	19.5
1	5.6	0	0	94.4	35.3	0.9	0.7	63.1	58.1	3.6	29.6	8.7
0	6.7	0	0	93.3	42.1	0.5	0.5	56.9	62.3	1.6	31.4	4.7

UC=UnderCover, OC=OverCover, F1=Fail, Gd=Good.

These simulations show that the performance of the  $H$  criterion is excellent with any value of  $\lambda$  when the target subset has a substantially larger proportion of successes compared to the rest. However, when the difference between the subset and the rest is small, setting  $\lambda$  to about 5 offers an improvement by forcing selection of subsets that are not too small. Setting  $\lambda$  much larger than 5 may result in picking intervals that are too big leading to overcovering; conversely, setting  $\lambda$  much smaller than 5 may result in picking intervals that are too small leading to undercovering. Similar patterns of results are obtained from simulations with different values of  $p(Y=1|I_T)$  and  $p(Y=1|D-I_T)$  and  $N$ . Based on these findings, we recommend setting  $\lambda$  to 5.

## 5.2 Performance

We carried out a simulation to assess the performance of ARF. Each simulation involved 100 runs, in which, for each run,  $D$  consisted of 2000 observations  $\{(Y_i, x_{i1}, x_{i2}, x_{i3})\}$ , with  $x_{ik}$  randomly selected from between -1 and 1. The target subset  $S_T$  was set to be the 25% subset  $\{(x_1, x_2, x_3): |x_1| < 0.5, |x_2| < 0.5\}$ . Simulations were run with various different values for  $p(Y=1|D)$  and  $p(Y=1|S_T)$ . The percentage *In* of observations in  $S_T$  that are correctly identified as being within  $S_T$  and the percentage *Out* of observations in the subset selected that are incorrectly identified as being within  $S_T$  are shown in Table 5 for a representative subset of simulations. In the null situation,  $p(Y=1|S_T)=p(Y=1|D)$ , *In* is small and *Out* is large, as expected. However, when  $p(Y=1|S_T)$  exceeds  $p(Y=1|D)$ , even by a moderate amount, *In* is large and *Out* is small, demonstrating ARF's effectiveness at finding HARs.

Table 5: Results of simulation to assess the performance of ARF.

$P(Y=1 D)=0.15$			$P(Y=1 D)=0.30$		
$P(Y=1 S_T)$	<i>In</i>	<i>Out</i>	$P(Y=1 S_T)$	<i>In</i>	<i>Out</i>
0.15	30.1	69.9	0.30	16.6	71.3
0.30	70.8	16.4	0.45	53.4	20.6
0.45	84.6	2.0	0.60	77.7	2.6

## 6. COMPUTATIONAL ALGORITHMS

The algorithm required to optimize the  $H$  criterion is a major issue due to the potentially large number of both observations and variables. We present here a series of algorithms that can be used to calculate the optimal interval or approximate optimal intervals for one continuous predictor variable  $X$ . The first algorithm calculates the global maximum of the criterion function  $H$  over all intervals. Later, we describe modifications that speed up the algorithm but produce a local maximum.

### 6.1 The main algorithm

We now describe how to determine the global maximum. We assume, without loss of generality, that  $H$  measures the proportion of  $Y=1$ 's in an interval determined by the values of  $X$ , plus a penalty for interval size. The naïve algorithm, which calculates the criterion  $H$  for all possible subsets of  $X$ , is computationally intensive. It requires evaluating the  $H$  criterion on  $N(N-1)/2$  intervals. Our algorithm works by dramatically reducing the number of intervals to be evaluated.

We consider  $\mathbf{y}$ , the sequence of  $Y$ 's (0s and 1s) sorted according to their corresponding  $X$  values and look at the intervals that begin with a 1 not preceded by a 1 and end with a 1 not followed by a 1. The algorithm proceeds as follows.

STEP 1: Set  $\mathbf{L}$ = list of indices in  $\mathbf{y}$  of 1s not preceded by a 1 and set  $\mathbf{U}$  = list of indices in  $\mathbf{y}$  of 1s not followed by a 1.

STEP 2: For all  $i$  and  $j$  such that  $U_j - L_i \geq M$ , (*i.e.*, for all intervals satisfying the minimum size constraint  $M$ ), evaluate the criterion function  $H$  in the interval defined by  $(L_i, U_j)$ .

STEP 3: Select the interval  $(L^*, U^*)$  that maximizes  $H$  and produce the corresponding  $(X_l, X_u)$  interval.

The final interval  $(X_l, X_u)$  is the global maximum for the criterion function  $T$ .

This algorithm is still of order  $O(N^2)$  but the number of evaluations of  $H$  is reduced enormously from the naïve algorithm. Even in the worst case, which occurs when  $\mathbf{y}$  is a sequence of alternating 0s and 1s, the numbers of  $L$ 's and  $U$ 's is  $N/2$ , which is half of the naïve algorithm, the computation is reduced by a fourth. In typical data mining problems, where a relatively small number of 1s (say 10%) could be expected, this algorithm would easily cut the computation by a factor of hundreds. This makes the computation feasible even for datasets with hundreds of thousands to a million cases on a fast personal computer.

*Remark:* When there are tied values of  $x$  associated with both 0s and 1s, the  $L$ s and  $U$ s of the ties are added to the list of  $L$ s and  $U$ s and anything in between is deleted.

## 6.2 Further reductions

If the reduction in number of computations described above is not sufficient, there is an alternative algorithm for reducing the computational time at the expense of calculating a good local maximum rather than a global one. The idea is to reduce the sequences  $L$  and  $U$  by preprocessing them. The reduction is based on the fact that the sequences  $L$  and  $U$  are intercalated, i.e., each  $L_i$  is followed by  $U_i$ , and then  $L_{i+1}$ , etc.. By definition, the indices between  $U_i$  and  $L_{i+1}$  correspond to 0s in the  $\mathbf{y}$  sequence. Let  $Z_i$  denote the number of 0s between  $U_i$  and  $L_{i+1}$ ;  $Z_i = L_{i+1} - U_i + 1$ . We eliminate  $U_i$  and  $L_{i+1}$  from their respective sequences when the number of 0s in between them is less than a constant, say  $Z_i \leq k$ .

STEP 1. Order the subgroups by the proportion of 1s, from the lowest to the highest, and calculate  $H$  for the cumulative subsets starting from the highest. At the top of the list we will have those categories that contain only 1s (if any), while at the bottom we will find those with only 0s (if any). Let  $a$  be the index of the first class containing some 0s and let  $b$  be the index of the last class containing some 1s.

STEP 2. Starting at the bottom,  $b$ , switch the  $i$ th and the  $(i+1)$ th positions and compare the value of  $H$  at the cumulative subset before and after the switch. If there is an increase in  $H$ , then keep the groups switched, otherwise undo the switch and go on to the next position, until the top,  $a$ , is reached.

STEP 3. Repeat this process until the sequence becomes stable, i.e., no switches improve  $H$ . Then, determine the subset that maximizes  $H$ .

We could also proceed in steps, i.e., start by eliminating all the  $Z_i \leq 2$ , then  $Z_i \leq 3$  and so on, until the sequences left in  $\mathbf{L}$  and  $\mathbf{U}$  have a length that makes the computation feasible. At this point, we would use the reduced  $\mathbf{L}$  and  $\mathbf{U}$  to calculate the optimal interval. Since the optimal interval will generally contain only a small subset of the data, we should be able to rerun the algorithm over the optimal subset using the full sequences  $\mathbf{L}$  and  $\mathbf{U}$ .

Another way of reducing the computations for the optimal ARF interval for very large datasets is to apply a binning algorithm. The start and end points of the bins are used to generate the sequences  $\mathbf{L}$  and  $\mathbf{U}$  and the same algorithm as above is applied. The resulting interval can be refined to determine a true local maximum.

### 6.3 Nominal categorical variables

A slight modification is necessary for nominal categorical predictor variables. Suppose that the variable (call it  $X$ ) takes one of  $m$  different values. Then there are  $2^m - 1$  possible subsets of  $Y$  generated by combinations of values of  $X$ .  $H$  can be evaluated at each subset and the optimal one selected. This is slow. A faster algorithm is as follows:

This algorithm has the same structure as a bubble sort.

### 6.4 Code

SPlus/R code that implements the ARF methodology is available on the website <http://www.rci.rutgers.edu/~cabrera/dm/dm.html>.

## 7. OTHER SITUATIONS

The  $H$  criterion and ARF procedures are very versatile and can be used in a variety of situations with reasonably straightforward modifications. This is demonstrated by several examples below.

The importance of a subset can be assessed as in Section 2.2. However, since the criterion is different, a new cutoff,  $h(N, \pi, \lambda)$ , would have to be used to judge whether  $H_{obs}$ , the observed value of  $H$ , is "significant". This can be done as described there. An alternative general procedure is to use rerandomization. This is done by rerandomizing the  $Y$ s, recalculating  $H$  for all intervals and finding the maximum  $H$ . This process is repeated several times. If the percentage of times in which the rerandomized value of  $H$  exceeds  $H_{obs}$  is low (less than 5% say), then the finding can be regarded "significant". When using  $H$  recursively, this rerandomization would have to be done at each step  $s$  to account for the different sizes and compositions of  $D_{s-1}$ .

### 7.1 Multiple Bernoulli responses

The  $H$  criterion can be extended to the case of multiple Bernoulli responses. Suppose that there are  $t$  of them with  $P(Y_{ij}=1|D)=p_j(D)$  for  $j=1, \dots, t$ . Letting  $P(Y_{ij}=1|I_k)=p_j(I_k)$ ,  $d=(p_1(I_k)-p_1(D), \dots, p_t(I_k)-p_t(D))$ ,  $V$ =covariance matrix of  $(Y_{i1}, \dots, Y_{it})$  and  $i(d)=1$  if  $d_j > 0$  for all  $j$ ,  $i(d)=0$  otherwise, the criterion is:

$$H = i(d)(d'V^{-1}d)^{1/t} + \lambda \log(h)/\log(N).$$

*Example:* In the SADB, to find an HSR for efficacy assay A1 and toxicity assay A2, we recoded assay A2 results so that  $Y=1$  implied absence of toxic activity. Then we ran ARF1 with the above  $H$  criterion with the assay A1 results as  $Y_1$  and the recoded assay A2 results as  $Y_2$ . The results of this analysis with  $\lambda=5$  (see Table 6) are an improvement to those obtained by using AS (as in Section 3, Table 2). The improvement occurs because here (1) fewer data points were discarded due to missing response values (2) the criterion takes the correlation between  $Y_1$  and  $Y_2$  into account.

Table 6: Result of running ARF1 to find an HSR.

Var	Min	Lower	Upper	Max	%1Y <sub>1</sub>	%1Y <sub>2</sub>	%Interval	N	Crit	Sf
IC50	0.07	3.29	41.8	61600	83.05	49.12	28.2	201	8.858324	S
ENERGY	13.33	44.39	147.37	147.53	91.42	54.29	38.3	77	5.784782	S

## 7.2 Comparing two treatments

When, as in Case Study 4, the objective is to track the characteristics that result in a difference in response rate between two groups, the  $H$  criterion can be modified in terms of a log-odds-ratio. If  $Y_g$  denotes the response in group  $g$ , for interval  $I_k$ , let  $n_{11}$ ,  $n_{12}$ ,  $n_{21}$ ,  $n_{22}$  be the entries in the  $2 \times 2$  contingency table formed by  $Y_1$  and  $Y_2$  within  $I_k$ . The log-odds-ratio is

$$\omega(I_k) = \log((n_{11}+0.5)(n_{12}+0.5)/(n_{21}+0.5)(n_{22}+0.5))$$

with asymptotic standard error estimated as

$$ASE(\omega) = \sqrt{(1/(n_{11}+0.5) + 1/(n_{12}+0.5) + 1/(n_{21}+0.5) + 1/(n_{22}+0.5))}$$

(Agresti (1990)). Since

$$z(I_k; D) = (\omega(I_k) - \omega(D)) / ASE(\omega)$$

is approximately standard normal, the  $H$  criterion is, as before,

$$H = z(I_k; D) + \lambda \log(h) / \log(N).$$

*Example:* For Case Study 4, the  $H$  criterion with  $\lambda=0$  finds that patients with baseline hemoglobin levels between 11.3 and 13.1 have a substantially higher response rate with treatment (94%) compared to placebo (20%); the response rates for the entire study were 83% (treatment) and 46% (placebo).

## 7.3 The continuous response case

When  $Y$  is a continuous response variable and it is of interest to determine subsets of high mean response values, the  $H$  criterion can be used after modification:

$$z(I_k; D) = \sqrt{n(I_k)} (y(I_k) - y(D)) / s(D),$$

$$H = z(I_k; D) + \lambda \log(h) / \log(N).$$

where  $y(D)$  and  $s(D)$  are respectively the mean and standard deviation of the  $\{Y_i\}$  in  $D$  and  $y(I_k)$  is the mean of the  $\{Y_i\}$  in the interval  $I_k$  under consideration.

*Example:* We ran ARF for Case Study 2 with the  $H$  criterion for continuous response (ave( $Y$ )=1.83,  $N$ =4703). The significant part of the output is shown in Table 7. The most interesting subset is defined by HIP96 (HIP96<768) and contains 10% of the data with a mean of 3.90. Within this there are three subregions, those defined by the nodes: CC (HIP96<768 and IR $\in$ [9457,25527]), CCC (HIP96 $\in$ [223,273] and IR $\in$ [9457,25527]) and CCLCC (HIP96 $\in$ [140,166], IR $\in$ [9457,25527] and OUTV<44520), that also look interesting.

Table 7: Result of running ARF2 for Case Study 2.

Node	Var	Min	Lower	Upper	Max	Mean	%Interv	Crit	N
C	HIP96	68	136	768	768	3.90	37.05	10.33	463
CC	IR	1414	9457	25527	36450	4.12	63.93	6.66	297
CCC	HIP96	136	223	273	768	4.99	14.14	6.36	43
CCL	HIP96	136	136	223	768	3.99	16.84	0.00	210
CCLC	OUTV	0	0	44520	1575371	4.55	37.14	6.84	79
CCLCC	HIP96	136	140	166	219	5.18	39.24	6.49	32

HIP96=number of hip replacements in 1996, OUTV=number of outpatient visits per year, IR=Inpatient revenue. See Table 2 for descriptions of the columns.

*Remark:* If the response is believed to contain outliers, then  $y(I_k)$ ,  $y(D)$  and  $s(D)$  should be replaced by robust analogs.

## 8. A BASIC PARADIGM OF DATA MINING

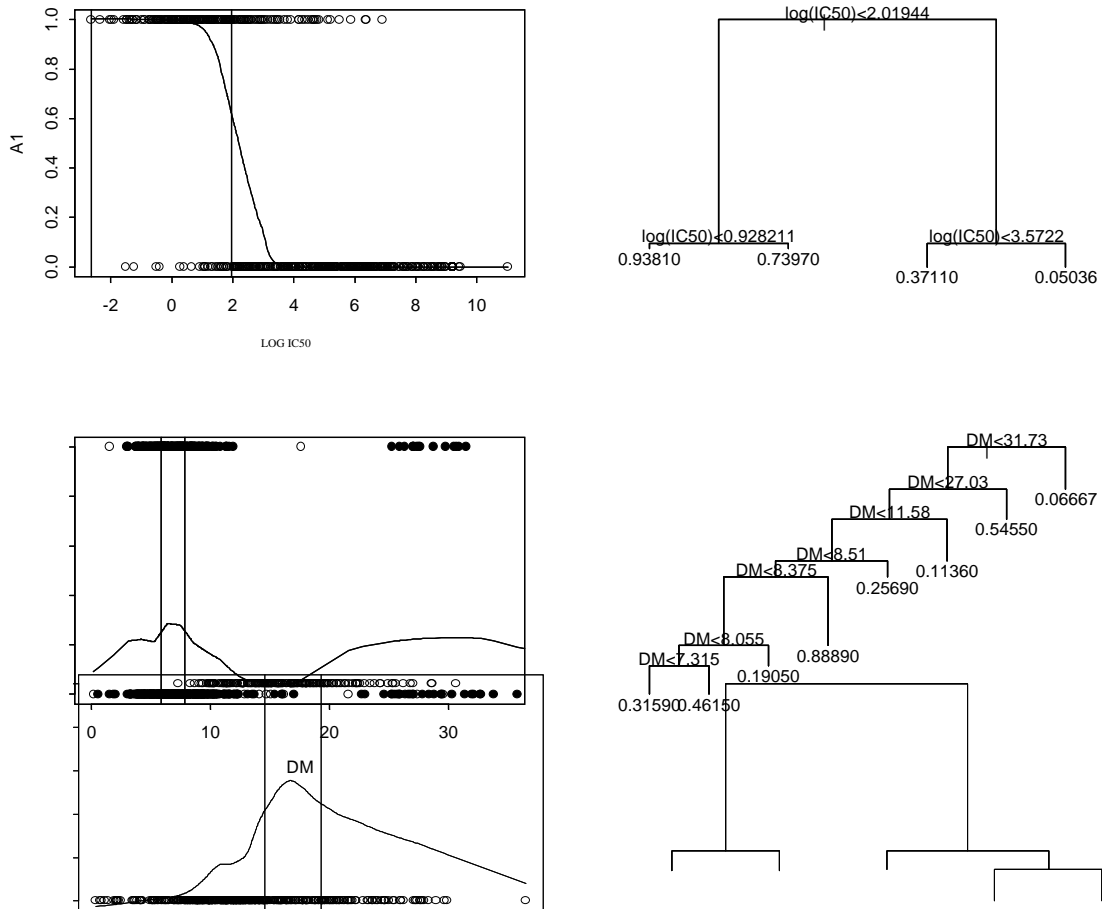
A pattern of interest in data mining can usually be thought of as a local optimum of some criterion function, which is comprised of the objective of interest together with a penalty for subset size, evaluated at the subset only. This can be regarded as a basic paradigm of data mining. ARF is one implementation of this paradigm. Further developments on this line, *e.g.*, for clustering (unsupervised learning), are in progress.

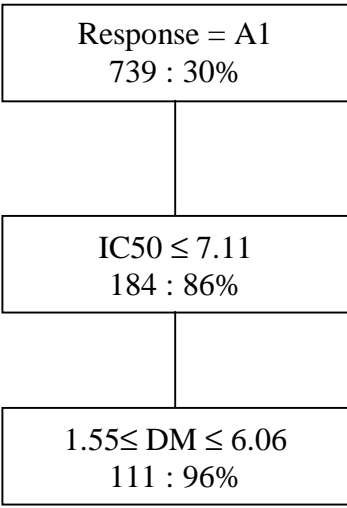
### REFERENCES

- Agresti, A. (1990), *Categorical Data Analysis*, New York, NY: John Wiley.
- Blake, C.L. & Merz, C.J. (1998), *UCI Repository of machine learning databases* [<http://www.ics.uci.edu/~mllearn/MLRepository.html>], Irvine, CA: University of California, Department of Information and Computer Science.
- Breiman, L., Friedman J.H., Olshen, R.A., and Stone, C.J. (1984), *Classification and Regression Trees*, Chapman and Hall.
- Clark, L.A. and Pregibon, D. (1992), Tree-based models, in *Statistical Models in S* (edited by J. Chambers and T.J. Hastie), Wadsworth.
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.
- Friedman, J.H. (1997), Data mining and statistics: what' s the connection?*unpublished*, <http://www-stat.stanford.edu/~jhf/>.
- Gordon, L., Schilling, M.S., and Waterman, M.S. (1986), An extreme value theory for long head runs, *Journal of Probability Theory and Related Fields*, 72, 279-287
- Hartigan, J.A. (1975), *Clustering algorithms*, New York, NY: John Wiley.
- Hawkins, D.M. and Kass, G.V. (1982), Automatic Interaction Detection, in *Topics in Multivariate Analysis* (edited by D.M.Hawkins), Cambridge University Press.
- Lee, Y.S. and Buja, A. (1999), Data mining criteria for tree-based regression and classification, *unpublished*, <http://www.research.att.com/~andreas/>.
- Morgan, J.N. and Sonquist, J.A. (1963) Problems in the analysis of survey data and a proposal, *Journal of the American Statistical Association*, 58:415-434.
- Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kauffman (<http://www.rulequest.com/see5-info.html>)
- Schilling, M.S. (1990), The longest run of heads, *The College Mathematics Journal*, 21:196-207
- Tukey, J.W. (1962), The future of data analysis, *Annals of Mathematical Statistics*, 33: 1-67.
- Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

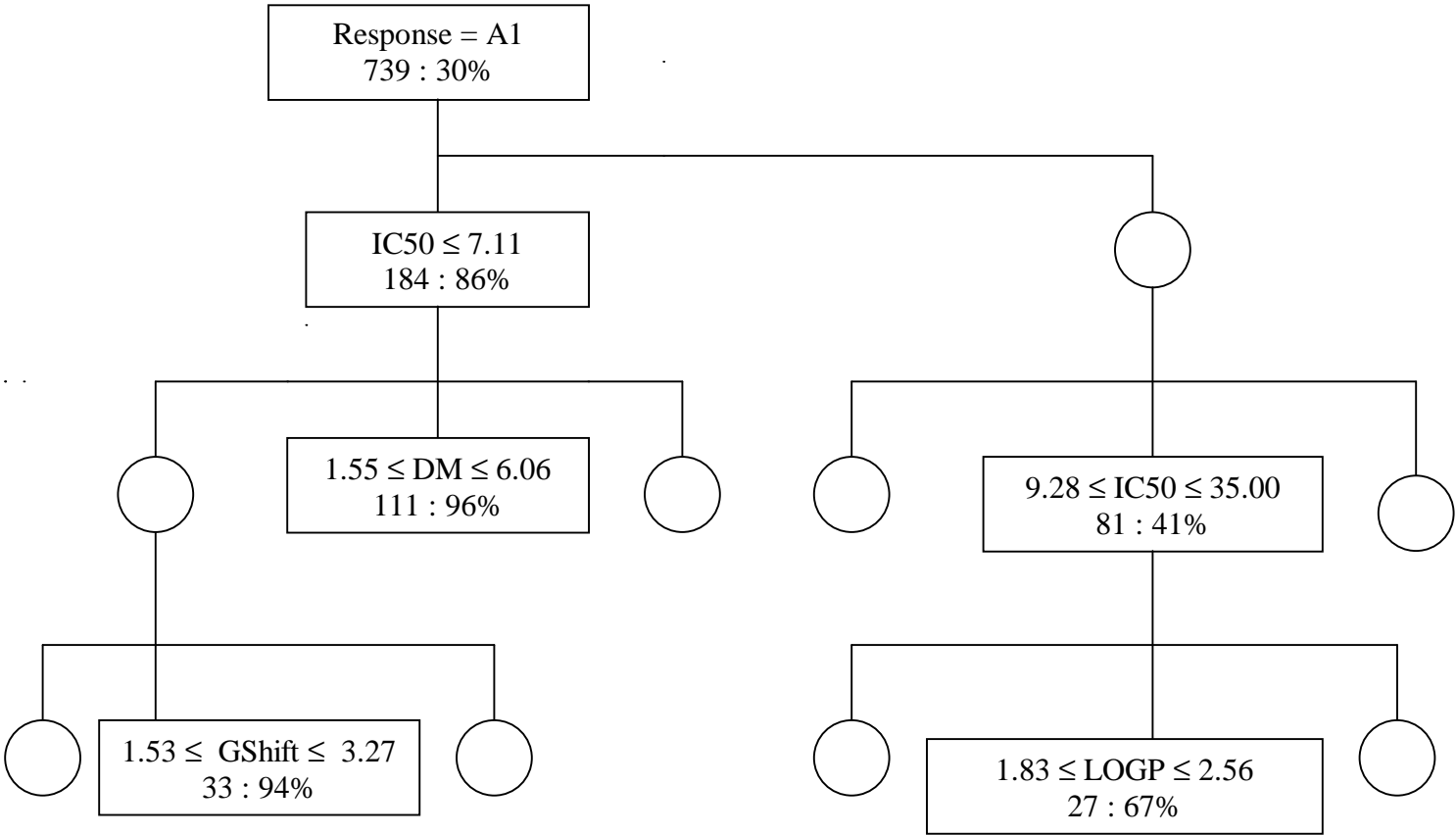
## LIST OF FIGURES

- Assay A1 results plotted against IC50, DM and MOLWT. A lowess smooth is overlaid on each plot. The vertical lines in the plots are the intervals that have the largest value of  $H$  for that variable. The column on the right gives the standard SPlus classification trees after pruning.
- (a) ARF1 tree and (b) ARF2 tree for assay A1 results. Significant nodes are in rectangles and nonsignificant nodes are in circles. The bottom line in each rectangle shows the number of observations in the interval and the success rate within it.





(a)



(b)