FAST-AND-FRUGAL TREES FOR DECISION-MAKING

Laura Martignon[1], Ulrich Hoffrage[2], Jan Woike[3], Tim Erickson[4] and Joachim Engel[1]
[1]Ludwigsburg University of Education, Ludwigsburg, Germany
[2]University of Lausanne, Lausanne, Switzerland
[3]MaxPlanck Institute for Human Development, Berlin, Germany
[4]Epistemological Engineering, Oakland, CA, USA
martignon@ph-ludwigsburg.de

Fast-and-frugal trees for classification/decision are at the intersection of three families of models: lexicographic, linear and tree-based. We briefly examine the classification performance of simple models when making inferences out of sample, in 11 medical data sets in terms of Receiver Operating Characteristics diagrams and predictive accuracy. The heuristic approaches, Naïve Bayes and fast-and-frugal trees, outperform models that are normatively optimal when fitting data. The success of fast-and-frugal trees lies in their ecological rationality: their construction exploits the structure of information in the data sets. The tool ARBOR, a digital learning tool, which is a plug-in to the freely available data-science education software CODAP can be used for constructing and interpreting fast-and-frugal classification and decision trees. This paper is an abridged version of work by Woike, Hoffrage & Martignon on the integration of classification and decision models into a common framework (Woike, Hoffrage & Martignon, 2017).

INTRODUCTION

Fast and frugal trees are models for classification and decision based on several pieces of information: Imagine a physician who wants to either confirm the presence of a particular disease given a set of observations or assure her patient that she does not need worry. In the Bayesian approach the physician begins with a prior probability, i.e., the base rate of the disease which is the probability P(D) that disease D is present. She then observes evidence E in the form of a set of cues (observations, symptoms, tests) and, based on these symptoms, assesses the probability that the evidence will occur if the disease is present and the probability that the evidence will occur if the disease is *not* present. For one piece of evidence *E* she computes *P(D|E)*, also called the posterior probability of the disease, given the evidence:

$$P(D \mid E) = \frac{P(E \mid D)P(D)}{P(E \mid D)P(D) + P(E \mid \overline{D})P(\overline{D})}$$

As a flurry of empirical findings have shown, this seemingly simple formula causes systematic errors when used by students of probability courses and even specialists (doctors, for instance). Changing the information format to so called natural frequencies has proven to greatly facilitate not just physicians' information processing but that of overall adults, as well as school students. Here the decision-maker imagines a group people with and without the disease, and then subdivides this population into subclasses (having or not having the disease), which are again subdivided according to the test results (positive and negative). Importantly the proportions of these subclasses correspond numerically to the original probabilities involved (observe that this is possible as long as probabilities are rational numbers). The necessary computation at the end involves simple arithmetic: the number of true positives divided by the sum of the number of true positives and the number of false positives. The beneficial effect of this format has been tested in a variety of applied domains, such as medicine, law, and education. It has furthermore been shown empirically that the beneficial effects of natural frequencies extends to more complex situations than those involving just one cue (symptom, test, or feature). Participants who know about the base rate or prior probability of a disease and have the statistical information about two independent tests or symptoms (specifically, their sensitivities and false-alarm rates) in terms of probabilities find it hard to infer the probability of the disease being present if both tests were positive. When they are provided with the corresponding information in terms of natural frequencies, about three quarters of the participants work out the correct (i.e., Bayesian) solution. Thus the facilitating effect of natural frequencies is generalizable to more cues.

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. ©2019 ISI/IASE iase-web.org/Conference_Proceedings.php

Nevertheless, there are obvious limits to the number of cues people can cope with. Making inferences under uncertainty even when provided with "natural" information formats becomes cumbersome when the number of cues exceeds what people can handle. One problem is limited memory, another is "brittleness or lack of robustness:" As the tree grows, the number of end nodes becomes ever larger, and the number of cases per end node becomes ever smaller. For some of the nodes the decision maker may not even have encountered specific cue–value combinations. Also, it is quite possible that for some specific combinations of cue values, the few cases encountered do not to generalize to new cases. It is here that heuristics become handy: The potentially huge full tree with natural frequencies and 2^{n+1} nodes when n is the number of cues can be pruned systematically into a minimal tree that may eventually come to use the same set of cues when necessary, often stopping after having checked just a few of them. This minimal tree turns out to be simple in construction, accurate in prediction and even robust for generalization.

These minimal trees have been called fast-and-frugal trees for classification and decision. By now, after their conceptual inception in 2003 (see Martignon et al., 2003), they have become popular even in the Machine Learning community. We have evidence that their properties and advantages can be taught to school students in ninth and tenth grade with success.

We begin by shortly presenting the most famous known example of fast-and-frugal tree, i.e., the one for assessing whether a patient arriving at the hospital suffering from severe chest pain is at high risk for myocardial infarction and should be assigned to the Coronary Unit or is moderately in danger and can be sent to a regular bed (the first such tree was described by Green and Mehr, still under the name of "Take The Best" in 1997). The data on 89 patients with severe chest pain were then collected in a Michigan Hospital and we use them here for modeling by means of two different trees in Figure 1: ST denotes a particular pattern of extreme elevation in the electro cardiogram, CP denotes chest pain, OS denotes "at least one other of 4 typical symptoms," "+" denotes present, and "–" denotes absent. Numbers in circles denote numbers of patients. Panel B: Fast-and-frugal classification tree obtained by pruning the natural frequency tree. Questions in rectangles specify which cues are consulted for each patient in the corresponding circle in Panel A. Depending on whether this cue value is positive or negative, either a new question is asked or the tree is exited and a decision is made (oval). The accuracy of these classification decisions is shown by the "patient numbers" below these oval exit nodes: The number of those who actually had a heart attack is displayed in a gray circle; the number of those who did not, in a white circle. All patients to the left of the vertical bar in Figure 1B are classified as high risk: all patients to the right, as low risk (Figure adapted from Woike, Hoffrage & Martignon, 2017).

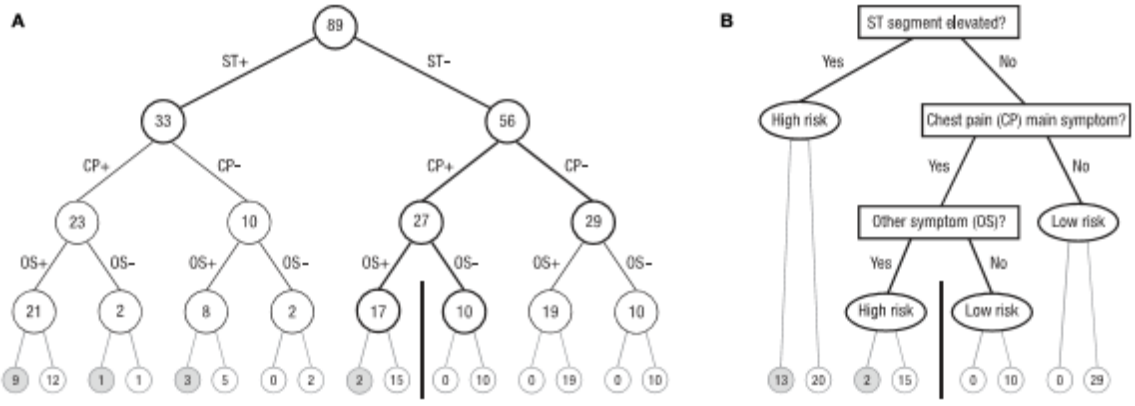

Figure 1: A full natural frequency tree (A) and the corresponding fast and frugal tree (B)

MATHEMATICAL PROPERTIES OF FAST AND FRUGAL TREES

The fast and frugal tree in Figure 1(B) presents a bold vertical bar that splits the cue profiles in two groups: those that are to the left of this bar and those that are to the right. The cue profiles to the left correspond to patients who are classified as being at high risk of a myocardial infarction.

Obviously patients with cue profiles [1, 1, 1], [1, 1, 0], [1, 0, 1], [1, 0, 0], or [0, 1, 1] are classified by the fast and frugal tree as being at high risk, and those with [0, 1, 0], [0, 0, 1], or [0, 0, 0]

as low risk. This splitting profile is a characteristic of fast and frugal trees. In fact they may be viewed as models that classify objects *lexicographically*. To explain this some mathematical rigor is required. We assume n that we work with n binary cues and two classes or categories, C_1 and C_0 (corresponding to the presence of disease and absence of disease, respectively). Without loss of generality, we also assume that cues are inspected in the order $c_1, c_2, \ldots, c_n$ and that for i = 1, 2, . . ., n, cues are coded so that in case an object x exits the tree at the i-th level, it is assigned to category C_1 if $x_i = 1$ and to C_0 if $x_i = 0$.

> **Definition**: A cue profile x is lexicographically larger than a cue profile y (x > y) if and only if there exists $1 \leq i \leq n$ such that $x_i = 1$ while $y_i = 0$ and $x_j = y_j$ for all $j < i$.
> If neither x > y nor x > y, then x and y obviously coincide.

The following result establishes a characterization of fast and frugal trees as lexicographic classifiers based on splitting profiles (Martignon et al., 2003; Martignon et al., 2008).

> **Result 1**: For every fast and frugal tree ***f*** there exists a unique cue profile S(f)—called the tree's splitting profile—so that ***f*** assigns x to C_1 if and only if x >l. Furthermore, for every cue profile S there exists a unique fast and frugal tree ***f***, such that S(***f***) = S (see Martignon et al. 2003).

In Figure 1(B), let $x_1 = 1$ if and only if the ST segment is elevated, $x_2 = 1$ if and only if chest pain is the main symptom, and $x_3 = 1$ if and only if any of the other four symptoms are present. Also let C_1 represent *high risk* and C_0 *low risk*. The splitting profile of this tree is [0, 1, 0]. The bold vertical bar marks the position of the splitting profile. All cue profiles to the left of the bar are lexicographically larger than the splitting profile. As result 1 states, these cue profiles are assigned to the high risk category C_1. The full "natural frequency tree" that adopts the same convention of placing positive (negative) cue branches to the left (right) does not have a splitting profile: If a threshold of .1 is adopted against which the posterior probability is compared, then the classifications for the eight distinct cue profiles are, from left to right, C_1, C_1, C_1, C_0, C_1, C_0, C_0, and C_0. For the fast and frugal tree, in contrast, the classifications for the four distinct cue profiles are, from left to right, C_1, C_1, C_0, C_0.

It can also be shown that a fast and frugal tree can be identified with a weighted linear model for classification with non-compensatory weights. In other words, fast and frugal trees can thus also be seen as belonging to the class of linear classifiers with non-compensatory weights. We recall that in linear models for classification, each cue c_i has a *weight* $w_i > 0$ and for each cue profile $x = [x_1, x_2, \ldots, x_n]$, the score $R(x) = \Sigma_i x_i w_i$ is computed. A scalar *threshold* $h > 0$ charactezises classification in the sense that item x is assigned to C_1 if and only if $R(x) > h$. A linear classifier in which all weights are 1 is called *Tallying*. The following result relates linear and lexicographic inferences for classifications (Martignon et al., 2008).

> **Result 2**: For every fast and frugal tree f there exist $h > 0$ and $w_i > 0$ where $w_i > \Sigma_{k>i} w_k$ for $i = 1, 2, \ldots, n-1$, so that f makes identical classifications with the linear model with weights w_i and threshold h. For every linear model with weights $w_i > 0$ so that $w_i > \Sigma_{k>i} w_k$ for $i = 1, 2, \ldots, n-1$ and a threshold $h > 0$, there exists a fast and frugal tree f that makes identical categorizations (the proof appears in Martignon et al., 2008).

For example, the Green and Mehr (1997) tree in Figure 1(B) makes identical classifications with the linear model with $R(x) = 4x_1 + 2x_2 + x_3$ and $h = 2$ (they both assign [0, 0, 0], [0, 0, 1] and [0, 1, 0] to C_0 and all other cue profiles to C_1). Result 2 states that fast and frugal trees are equivalent to non-compensatory linear models in the sense that the two make the same classifications. Note, however, that Result 2 does not imply that it is impossible to distinguish empirically between fast and frugal trees and noncompensatory linear models.

We will show how a fast and frugal tree, like the one displayed in Figure 1(B), may be constructed. The two important features of a fast and frugal tree are its *ordering of cues* and its *exit structure*. Keeping in mind that n cues allow for n! different orderings, the task of selecting useful orderings is not trivial. In fact, the task of finding the ordering that leads to best performance is NP-

complete (Schmitt & Martignon, 2006). Implementing orderings and searching for those with good performance would be too complex and far away from the spirit of a *simple heuristics program* which we propagate here. We propose heuristic principles for constructing trees.

In our approach, the ordering of cues is determined by their relationships with the criterion. Fast and frugal trees assume cues to be independent of each other, conditioned on the criterion. They are "naïve" in the same sense in which "Naïve Bayes Networks" are naïve.

For just one cue, say in the specific medical example above, its Positive Predictive Value (PPV) is the proportion of patients suffering from the disease among all patients for whom the cue is positive (or in its high state, i.e., p(D|E)), and the negative predictive value (NPV) is the proportion of patients without the disease among all patients for whom the cue is negative (or in its low state, i.e., p(–D|–E)).

In general, PPV and NPV indicate how diagnostic a cue is given that it has a positive or negative value, respectively.

HOW DO FAST AND FRUGAL TREES PERFORM?

Three large classes of classification strategies have been extensively evaluated in one common framework (Woike, Hoffrage & Martignon, 2017): full natural frequency tree, naïve Bayes and several variants of fast and frugal trees. The evaluation methods are, on the one hand, computations of the average between the PPV and NPV of a strategy, that is the overall diagnosticity of the strategy, and, on the other, the well-known ROC curves for measuring the *sensitivity* of a strategy, based on Receiver Operating Characteristics diagrams. These curves plot sensitivity (percentage of correct identification among patients with the disease) against false alarm rate (incorrect identifications among patients with the disease; the complement of the specificity) as the classification threshold is gliding from 0 to 1. Fast and frugal trees do not use such thresholds and hence their classification performance cannot be represented as a curve, but just by one point per construction principle. Each data point in the ROC diagram reveals these two proportions, but it is mute about the underlying absolute frequencies. The same data point can signal a high or a low accuracy depending on which of the two errors occurs how often. In other words, ROC curves display "normalized frequencies", rather than natural frequencies. We now list the different versions of very simple fast and frugal trees:

Rakes (R). Following Martignon et al. (2003) we will refer to a tree that has all exits to the same side as a rake. The tree that orders cues by PPV and that has all exit nodes on the left will be abbreviated by R+, and the one that uses NPV and has all exit nodes on the right by R-.

Zigzag trees (Z). We define Z+ as starting with the cue that has the highest PPV among all cues and placing it at the top of the diagnostic tree, with the exit node on its left. Subsequently, Z+ identifies the cue with the highest NPV among all remaining cues, places it second and with the exit node on the right. Then it identifies, for the third position, among the remaining cues the one with the highest PPV (and puts its exit node on the left), and so on. Z- follows the same logic but it starts with the cue that has the highest NPV, continues with the cue that has the highest PPV, then again the one with the highest NPV, and so on.

Base-rate respect trees (B). We now introduce trees that combine subtrees of the rake and of the zigzag type. It is easy to see how zigzag trees suffer from this base-rate neglect. Therefore we propose two trees that combine features of rakes and zigzag trees. In the standard variant, denoted as B, for a base rate of b<0.5 the first $k = \lfloor |\log_2(b)| \rfloor$ cues have exit nodes on the right, thereby classifying patients into the majority category. These k cues, and their ordering, are determined by maximum NPV. The B_1 tree is even more biased: It does not start with k, but with $k + 1$ cues that classify objects into the majority category.

Maximum predictive value trees (M). Our next tree is also quite sensitive to the environment. Like for all trees, the position of the exit is determined by whether the cue on this position has been picked by its PPV or by its NPV. The tree that is here referred to as M chooses the maximum predictive value among all remaining cues.

Accuracy trees (A). The PPV and the NPV are still considered, not to determine the position in the cue ordering, but only to determine the exit node. If for a given cue PPV>NPV, the exit is to the left, otherwise to the right. For each of the tree construction principles proposed above, ties in the process are broken randomly.

The data sets for our competition are listed in the following table:

Table 1. Data sets investigated

Panel in Figure 3	Name	Criterion	Base rate of positive criterion	Objects	Cues
A	Alcohol	>2.5 pints per day consumption	0.255	345	5
B	Echocardiogram	Survival after heart infarction	0.290	62	7
C	Diabetes	Diabetes	0.349	768	6
D	Breast Cancer	Malignant tumors	0.350	683	9
E	Heart Disease Hungarian	Heart disease	0.374	262	10
F	Heart Disease Cleveland	Heart disease	0.461	297	13
G	Horse Colic	Treated with surgery (yes/no)	0.587	218	6
H	Post-Operative	Decision to keep patient in hospital	0.721	86	8
I	Hepatitis	Survival of hepatitis patients	0.791	153	7
J	SPECT	Heart disease	0.794	267	22
K	Cardiac*	Heart disease	0.841	558	8

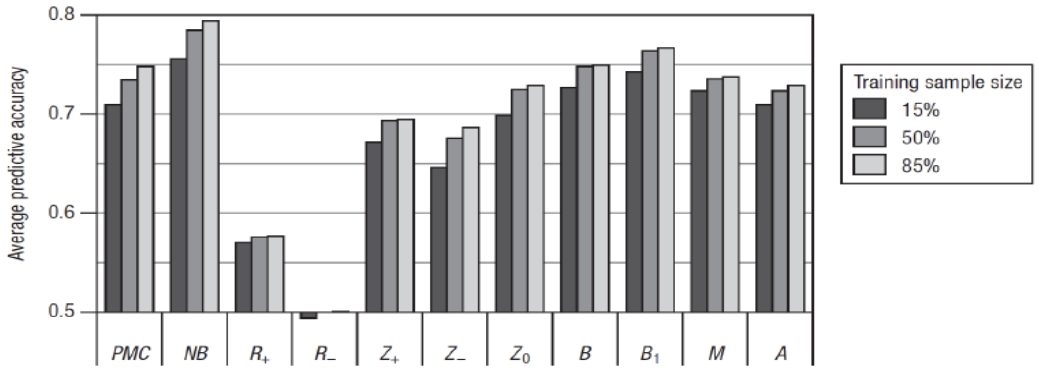

Figure 2: Table with our data sets and results: Predictive accuracy for profile memorization classification, Naïve Bayes, and various fast-and-frugal trees across all data sets, with size of training sample of 15%, 50%, and 85%. The performance of the classifier was evaluated in the remaining 85%, 50%, and 15% of cases.

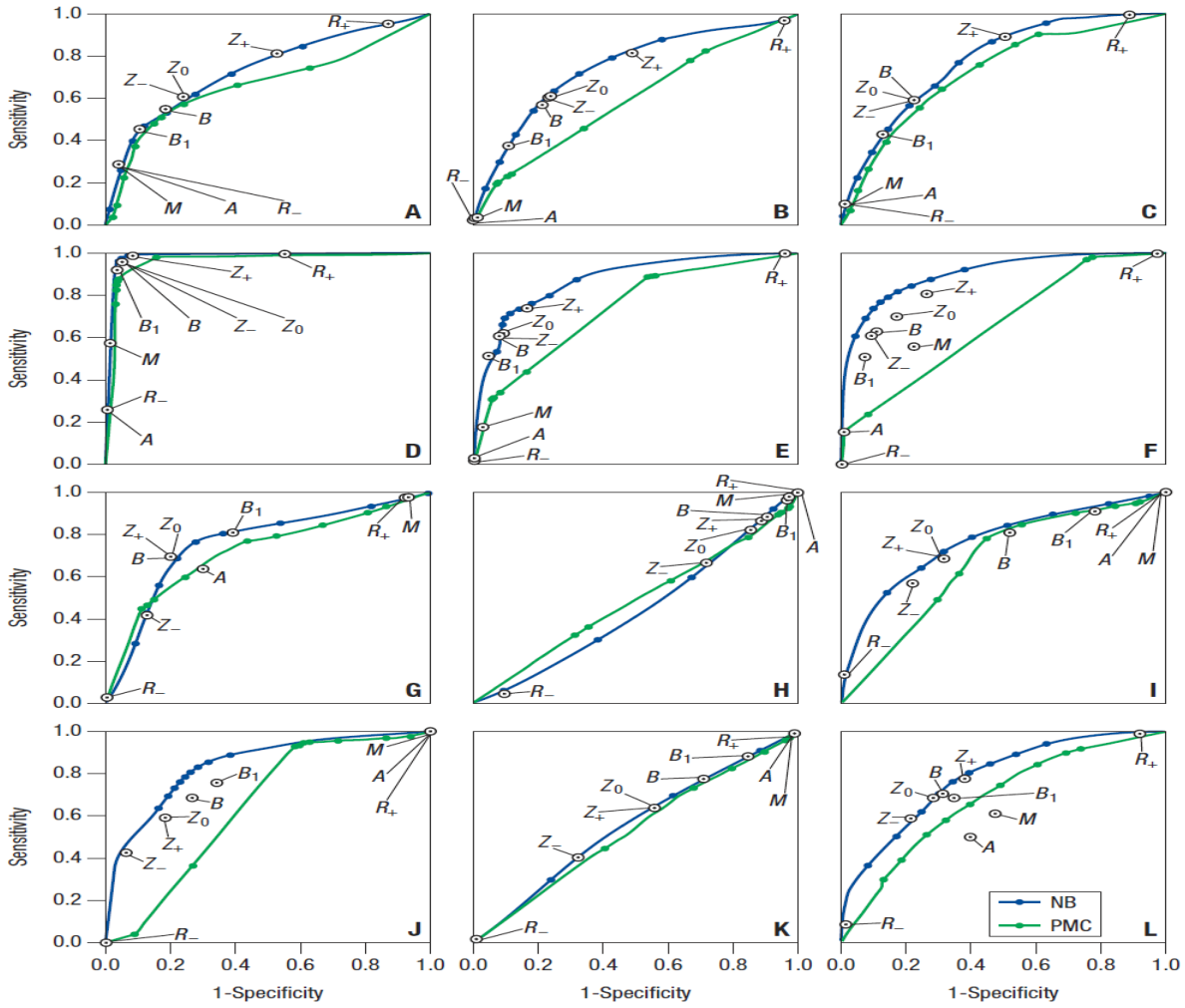

Figure 3: ROC Curves for profile memorization classification and Naïve Bayes, and the corresponding data points for various fast-and-frugal trees, separately for 11 data sets (Panels A–K) and averaged across all data sets (Panel L), with a training sample of 50% of the objects and performance tested on the remaining 50%. As we observe, the overall winners are Naïve Bayes and B_1 demonstrating that naïve heuristics do not trade off accuracy for simplicity. They are accurate not inspite of their simplicity but because of it!

For the possibility of visualizing the construction of fast and frugal tree by means of ARBOR the reader can consult the paper by Engel, Erickson & Martignon in this volume.

REFERENCES

Engel, J., Erickson, T., & Martignon, L. (this volume). Teaching about tree based decision-making.

Green, L., & Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *The Journal of Family Practice*, *45*(3), 219–226.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources. A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. (2003). Naïve and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman, & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 189–211). Chichester, United Kingdom: John Wiley and Sons.

Schmitt, M., & Martignon, L. (2006). On the complexity of learning lexicographic strategies. *Journal of Machine Learning Research*, *7*, 55–83.

Woike, J., Hoffrage, U., & Martignon, L. (2017). Integrating and testing Frequencies, Naïve Bayes and fast-and-frugal trees. *Decision, 4*(4), 234–260.