Proceedings of Machine Learning Research 58 (2016) 75-85

On Learning Decision Heuristics

Özgür Şimşek OZGUR@MPIB-BERLIN.MPG.DE
Marcus Buckmann BUCKMANN@MPIB-BERLIN.MPG.DE
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Editor: Tatiana V. Guy, Miroslav Kárný, David Rios-Insua, David H. Wolpert

Abstract

Decision heuristics are simple models of human and animal decision making that use few pieces of information and combine the pieces in simple ways, for example, by giving them equal weight or by considering them sequentially. We examine how decision heuristics can be learned—and modified—as additional training examples become available. In particular, we examine how additional training examples change the variance in parameter estimates of the heuristic. Our analysis suggests new decision heuristics, including a family of heuristics that generalizes two well-known families: lexicographic heuristics and tallying. We evaluate the empirical performance of these heuristics in a large, diverse collection of data sets.

1. Introduction

Decision heuristics (Gigerenzer et al., 1999, 2011) are models of human and animal decision making. They use few pieces of information and combine the pieces in simple ways, for example, by giving them equal weight or by considering them sequentially. We examine how such heuristics can be learned from training examples. Our motivation is both computational and cognitive. From a computational viewpoint, we want to develop heuristics that are fast in computation, frugal in information use, and effective in making good decisions. From a cognitive viewpoint, we want to understand how people and animals create and modify decision heuristics over time, as they accumulate experiences. While there is a large literature on decision heuristics, few studies have examined the learning process.

Simple decision heuristics are most widely studied within the context of comparison problems, where the objective is to identify which of a number of alternatives has the highest value in a specified (unobserved) criterion. We study the learning process in this context, examining two well-known families of heuristics: lexicographic heuristics and tallying. In particular, we examine the variance resulting from small training-set sizes and its effect on the building blocks of these heuristics.

Our theoretical analysis suggests new heuristics, including a generalization of lexicographic heuristics and tallying. We examine the performance of these heuristics in a large, diverse collection of data sets, comparing their performance to random forests. We find that sampling variance can have a large impact on the learning rate of heuristics and that very simple methods for accounting for sampling variance can substantially improve predictive accuracy.

2. Background

The comparison problem asks which of a given set of objects has the highest value on an unobserved criterion, given a number of attributes of the objects. We focus on pairwise comparisons, where exactly two objects are being compared. In the heuristics literature, *attributes* are called *cues*; we will follow this custom. We use x_A and y_A to denote the cue and criterion value, respectively, of object A.

For the comparison problem, two well-known families of heuristics are lexicographic heuristics and tallying. Both families decide by comparing the objects on one or more cues, asking which object has the more favorable cue value. Each cue is associated with a direction of inference, known as the *cue direction*, which can be positive or negative, favoring the object with the higher or lower cue value, respectively. Neither family requires the difference in cue values to be quantified. For example, if *height of a person* is a cue, one needs to be able to determine which of two people is taller but it is not necessary to know the height of either person or the magnitude of the difference.

Lexicographic heuristics (Fishburn, 1974) consider the cues one at a time, in a specified order, until they find a cue that *discriminates* between the objects, that is, one whose value differs on the two objects. The heuristic then decides based on that cue alone. An example is take-the-best (TTB; Gigerenzer and Goldsein, 1996), which orders cues with respect to decreasing validity on the training sample, where *validity* is the accuracy of the cue among pairwise comparisons on which the cue discriminates between the objects.

Tallying (Czerlinski et al., 1999) is a voting model. It determines how each cue votes on its own (selecting one or the other object or abstaining from voting) and selects the object with the highest number of votes, breaking ties randomly. Cue directions are set to the direction with the highest validity in the training set.

Note that the comparison problem has a symmetry: a comparison of A to B and B to A should agree on which object has the higher criterion value.

3. Distribution of sample statistics

A primary building block of decision heuristics is how a cue decides on its own, independently of the other cues. This is determined by the cue direction. If the direction is positive, the cue favors the alternative with the higher cue value; if it is negative, the cue favors the alternative with the lower cue value.

A second building block is how well a cue decides on its own, in other words, how accurate it is when it discriminates among the alternatives. This building block informs how the various cues should be integrated within the heuristic, for example, how they should be ordered in a lexicographic decision rule. For this building block, two quantities are relevant: the positive and negative validity, which are the probability that the cue makes the correct decision given that the cue discriminates between the alternatives if the cue is used in the positive or negative direction, respectively. *Cue validity* is the larger of positive and negative validity—it is the accuracy of the cue when it discriminates between the alternatives if the cue is used in the correct direction.

In earlier work (Şimşek and Buckmann, 2015), these building blocks were examined with a focus on expected rate of learning. Here, our main focus is on sampling variance in cue parameters.

When learning decision heuristics, cue directions and cue validities are estimated from a training sample, where each training instance corresponds to a single pairwise comparison between two objects. We assume that the instances in the training sample are independent.

From a comparison of object A to object B, the information we need is a single variable with three possible values: *positive* if the cue and the criterion move in the same direction, that is, if $(x_A - x_B) \times (y_A - y_B) > 0$; *negative* if the cue and the criterion move in opposite directions, that is, if $(x_A - x_B) \times (y_A - y_B) < 0$; and *neutral* otherwise. We can therefore denote a training sample with three numbers, $\{a, b, c\}$, where a is the number of positive instances, b the negative instances, and c the neutral instances.

Given a training sample, the estimate of cue direction, $\hat{d}$, is positive if $a > b$, negative if $a < b$, and positive or negative with equal probability if $a = b$. The estimate of cue validity, $\hat{v}$, is $\max\{a, b\}/(a + b)$. Notice that the value of c does not play a role in these estimates. Our analysis therefore focuses on samples with no neutral instances, where the sample size is $n = a + b$. We call n the number of *informative* instances. We denote the true validity and direction of the cue with v and d, respectively. In the analysis that follows, we assume, without loss of generality, that d is positive.

We examine the sample distributions of three variables. The first is $\hat{d}$, which is 1 if the cue-direction estimate from the sample is identical to the cue direction in the population, and 0 otherwise. The second variable is $\hat{v}$, and the third is o, which is the expected accuracy of the cue on an unseen test instance where the cue discriminates between the alternatives if the cue is used in the direction inferred from the sample. Our main objective in this section is to examine the variance in $\hat{d}$, $\hat{v}$, and o.

Lemma 1 Random variable $\hat{d}$ follows a Bernoulli distribution with probability of success $p_1 = \sum_{k=\lfloor n/2 \rfloor + 1}^{n} B(k, n, v) + 0.5 \times B(n/2, n, v)$, expected value p_1, and variance $p_1(1 - p_1)$, where $B(x, n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$ denotes the binomial function.

Lemma 2 Random variable $\hat{v}$ has expected value v and variance $v(1 - v)/n$. This follows from $\hat{v} = a/n$ and the fact that a follows the binomial distribution with parameters n and v, with expected value nv and variance $nv(1 - v)$.

Lemma 3 Random variable o has expected value $p_1(2v - 1) + 1 - v$ and variance $p_1(1 - p_1)(2v - 1)^2$. Proof: $o = \hat{d}v + (1 - \hat{d})(1 - v) = \hat{d}(2v - 1) + 1 - v$. It follows that $E(o) = E(\hat{d})(2v - 1) + 1 - v$ and $Var(o) = Var(\hat{d})(2v - 1)^2$.

First, we briefly examine the expected prediction error (E) of a single cue:

$$\begin{aligned} E &= (1 - v) \times P(\hat{d} = 1) + v \times (1 - P(\hat{d} = 1)) \\ &= (1 - v)p_1 + v(1 - p_1) \\ &= \underbrace{(1 - v)}_{irreducible} + \underbrace{(2v - 1)(1 - p_1)}_{reducible} \end{aligned} \tag{1}$$

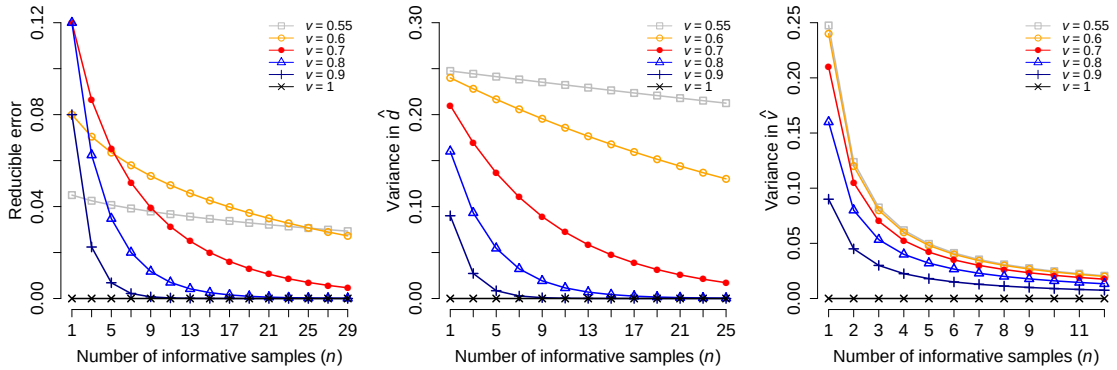

Figure 1: Reducible error, variance in $\hat{d}$, and variance in $\hat{v}$, as the number of informative samples increases.

The first term in Equation 1 is the irreducible error of the cue: This is the error that would be incurred even if the direction was known. The second term results from misestimation of the cue direction and becomes zero as sample size goes to infinity. Figure 1 (leftmost panel) shows the reducible error as a function of n for various v values. The plot shows an interesting pattern: When reducible error is high, it reduces rapidly; when it is low, it reduces very slowly. This conclusion was noted earlier (Şimşek and Buckmann, 2015); the analysis here is an alternative that shows it more clearly.

Figure 1 (middle and rightmost panel) shows how the variance in $\hat{d}$ and $\hat{v}$ decreases as sample size increases. Variance in sample validity reduces fairly rapidly within the first few samples. On the other hand, for cue direction, reduction in variance can be rapid or slow, depending on population validity. For high population validity, variance reduces rapidly. The closer the population validity is to 0.5, the slower the reduction in variance.

These results show that the sample variance of the cue-direction estimate varies substantially with population validity and with the number of informative samples. This has important consequences for learning heuristics. Even when the data set is complete, with no missing cue or criterion values, typically cues will vary (sometimes substantially) in the number of informative samples they have. For example, cues with lower discrimination rates will typically have smaller n.

Existing decision heuristics do not take this information into account. For example, tallying collects votes from all available cues, with no regard for how much uncertainty there is around the cue direction estimates. Similarly, TTB orders cues with respect to their sample validity, with no regard for the uncertainty around cue direction and validity estimates.

When the training set is large enough, the uncertainty in parameter estimates will diminish and not play a role, but with small training sets, there will often be substantial differences in how certain one is about the true direction and validity of the various available cues.

What can be done? In the next section, we turn our attention to the reverse inference problem: Given sample statistics, what is the true cue direction and validity in the population?

4. Inference on population statistics from a sample

We now examine how to make inferences on cue direction, cue validity, and o, given a training set $\{a, b, c\}$. First, we derive the posterior probability that population direction is positive given the training sample, $P(d_+|a, b, c)$. Note that the quantity c is irrelevant and $P(d_+|a, b, c) = P(d_+|a, b)$.

$$\begin{aligned} P(d_+|a,b) &= \frac{P(a,b|d_+)}{P(a,b)}\,P(d_+) = \frac{\int_{v=0}^{1} P(a,b|v,d_+)P(v|d_+)dv}{\int_{v=0}^{1} P(a,b|v)P(v)dv}\,P(d_+) \\ &= \frac{\int_{v=0.5}^{1} \binom{a+b}{a} v^a(1-v)^b\, 2\, P(v)\, dv}{\int_{v=0}^{1} \binom{a+b}{a} v^a(1-v)^b P(v)\, dv} 0.5 \\ &= \frac{\int_{v=0.5}^{1} v^a(1-v)^b P(v)\, dv}{\int_{v=0}^{1} v^a(1-v)^b P(v)\, dv} \end{aligned} \tag{2}$$

To arrive at Equation 2, we first used Bayes's rule, then conditioned on the population validity v, both in the numerator and in the denominator. Due to the symmetry of the comparison problem, $P(d_+)$ (prior probability of positive cue direction) is 0.5, and $P(v|d_+)$ is $2 \times P(v)$ if $v \geq 0.5$ and 0 otherwise.

Equation 2 uses $P(v)$, the prior on v. Figure 2 shows the distribution of cue validities in a large, diverse collection of natural data sets (described in section 6). The triangular distribution matches the validity distribution well.

Next, we derive the posterior distribution of population validity given sample statistics:

$$\begin{aligned} P(v|a,b) &= \frac{P(a,b|v)}{P(a,b)}\,P(v) = \frac{P(a,b|v)}{\int_{v=0}^{1} P(a,b|v)P(v)dv}\,P(v) \\ &= \frac{\binom{a+b}{a} v^a(1-v)^b}{\int_{v=0}^{1} \binom{a+b}{a} v^a(1-v)^b P(v)\, dv} P(v) \\ &= \frac{v^a(1-v)^b}{\int_{v=0}^{1} v^a(1-v)^b P(v)\, dv} P(v) \end{aligned} \tag{3}$$

To arrive at Equation 3, we first used Bayes's rule, then conditioned on the population validity v in the denominator.

And finally, we derive the posterior distribution of o, the probability that the cue will be accurate on a test instance, given that it discriminates between the alternatives, if the cue is used in the direction inferred from the training set.

$$\begin{aligned} P(o|a,b) &= \int_{v=0}^{1} P(o|v,a,b)P(v|a,b)\, dv \\ &= \int_{v=0}^{1} (v\hat{d} + (1-v)(1-\hat{d}))P(v|a,b)\, dv \end{aligned} \tag{4}$$

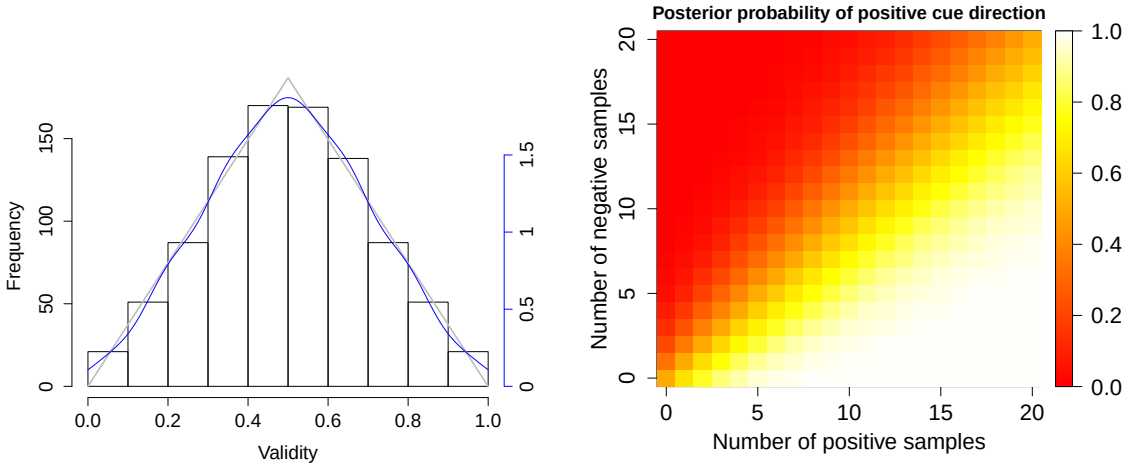

Figure 2: Left: A histogram showing the distribution of cue validity in 56 natural data sets. The blue line is a density estimate with a corresponding vertical axis to the right of the curve. The gray line is a triangular distribution that is a very close fit to the density. Right: The posterior probability that the population direction is positive given the number of positive and negative instances in the training set, computed by setting the prior on population validity to the triangular distribution shown on the left.

We now examine some of these quantities using the triangular distribution in Figure 2 as a prior on cue validity. Figure 2 (right panel) shows the posterior probability of positive cue direction as a function of a and b. The plot has a simple structure. There is a narrow band around the line $a = b$ where each additional informative sample changes the posterior substantially. For example, at (a = 3, b = 1), the posterior probability is 0.62, and after sampling one more positive instance, the posterior becomes 0.73. This band slowly expands in width as a and b increase.

The figure also shows that values of a and b that yield identical sample validities can have vastly different posterior probabilities of positive cue direction. Consider having four cues with training sets $\{2, 1\}$, $\{4, 2\}$, $\{6, 3\}$, and $\{8, 4\}$. All cues have sample validity 2/3 but their posterior probabilities of positive cue direction are 0.69, 0.78, 0.83, and 0.87, respectively. The general pattern is similar for the posterior probability that $o = 1$ (plot not shown).

It is important to note that while the computation of the posteriors is complex, the resulting posterior distributions exhibit simple patterns that simple rules for handling uncertainty would be able to express. This analysis shows that in early stages of learning, there are strong reasons to pay attention to the differences among cues in the level of uncertainty about their key parameters. In the next section, we consider three decision models that are sensitive to sample variance.

5. Decision models to consider

Motivated by the analysis in the previous sections, we consider three decision models. Two of the models are two extremes of the lexicographic decision rule where the training method ignores the dependencies among the cues. The first model answers, in a principled way, how to best take into account the differences in sample variances of the different cues. The second model asks to what extent this problem can be addressed using minimum effort and computation. As training-set size grows, both models converge to TTB. The third model takes a different perspective, and asks, if there is uncertainty about the order of the cues, why should one order the cues at all?

Lexicographic-by-posterior is a lexicographic decision rule that orders cues with respect to the posterior probability of a correct decision ($o = 1$), as computed by Equation 4, breaking ties randomly. This method is not computationally simple but represents an ideal for lexicographic decision rules that ignore the dependencies between the cues—it is the best that can be done. We call this method *lexipost* for short.

TTBS is a variation on TTB. It orders cues in decreasing order of sample validity (as TTB does) but breaks ties in favor of the cue with the higher number of informative samples. Variance in sample estimates of cue direction and cue validity reduces with increasing number of informative samples, and TTBS is one of the simplest, most straightforward methods of being sensitive to sample variance.

Lexicographic-tallying is a family of heuristics that generalizes lexicographic models and tallying. It is characterized by cue directions and cue levels. At decision time, at first, only the cues at level 1 are examined. These cues vote independently, and their independent decisions are tallied. If the result favors one or the other alternative, a decision is reached. Otherwise (if the tally at level 1 is neutral between the two alternatives), the cues at level 2 are tallied, and so on, until a decision is reached. If all cues are at level 1, the method reduces to tallying. If no cues share the same level, it reduces to a lexicographic decision rule. We refer to this model as *lexital* for short. We are not aware of earlier uses of this model even though it is a natural generalization of existing heuristics.

One motivation for using lexital is uncertainty in cue parameters. When there is not enough certainty about how a subset of the cues should be ordered (e.g., if they have equal sample validity and an equal number of informative samples), tallying these cues is a more reasonable approach than using them sequentially. There are other reasons for employing the hierarchical structure of lexital but uncertainty in cue parameters is a natural reason for doing so.

How should the parameters of a lexital model be determined from training samples? This is an open question, with many possible approaches. Here we consider perhaps the simplest approach: Order cues according to decreasing sample validity; when multiple cues tie on their sample validity, assign them to the same level in lexital. We call this model *tally-the-best*.

6. Empirical performance

We examined the performance of various heuristics in a large, diverse collection of natural data sets. The collection included 56 data sets gathered from a wide variety of sources,

including online data repositories, textbooks, packages for R statistical software, statistics and data-mining competitions, research publications, and individual scientists collecting field data. The subjects were diverse, including biology, business, computer science, ecology, economics, education, engineering, environmental science, medicine, political science, psychology, sociology, sports, and transportation. The data sets varied in size, ranging from 13 to 601 objects. Many of the smaller data sets contained the entirety of the population of objects, for example, all 29 islands in the Galápagos archipelago. Most of the data sets were used in earlier studies (Czerlinski et al., 1999; Şimşek, 2013; and Şimşek and Buckmann, 2015). All are publicly available. The data sets are described in the supplementary material.

We tested the following models: TTB, TTBS, lexipost, and tally-the-best. In addition, we tested random forests (Breiman, 2001), one of the very best statistical learning algorithms, to provide a strong benchmark from machine learning. We trained random forests using their implementation in R package *randomForest* (Liaw and Wiener, 2002). Typically, the only parameter tuned when using random forests is *mtry*, which specifies how many cues should be randomly selected for consideration when splitting a branch (Hastie et al., 2009). We tuned *mtry* using 10-fold cross-validation in the training set. A description of our random-forest implementation is provided in the supplementary material.

We focused on cases where sample variance plays a role in the learning process, for example, due to differences in discrimination rates, resulting in differences in the number of informative samples available for different cues (even though all cues were trained on the same set of paired comparisons). We observed (not so small) differences in the learning curves of TTB and lexipost in 25 of the 56 data sets.

Figure 3 (top left) shows the mean accuracy in these 25 data sets as the training-set size grows, starting with one instance. We focus here on differences in early stages of learning. Because some of the data sets are smaller than others, the number of data sets included in the figure decreases as training-set sizes increases (therefore the tail end of the learning curves are not smooth). On the 25 data sets, there is a substantial gap between TTB and random forest. TTBS, lexipost, and tally-the-best close this gap to some extent.

In the heuristics literature, it is common to dichotomize the cues around the median (Czerlinski et al., 1999; Brighton, 2006; Martignon et al., 2008), for which one reason is "to mimic the limited knowledge about cue values that people typically have, and the potential unreliability of precise values" (Gigerenzer et al., 1999). With dichotomized cues, sampling variance almost always plays an important role in the learning process. In Figure 3 (top right), we show mean accuracy in all 56 data sets when the cues were dichotomized around the median. All three models performed better than or as well as random forests. TTB lagged behind in some regions of the learning curve.

The figure shows, in addition, individual learning curves in 9 of the 25 data sets. These plots show two standard errors around each learning curve as a shaded region surrounding the curve above and below. On individual data sets, lexipost frequently made large improvements in performance compared to TTB. Surprisingly, this was also true for the other two methods, tally-the-best and TTBS, despite their very simple handling of sampling uncertainty.

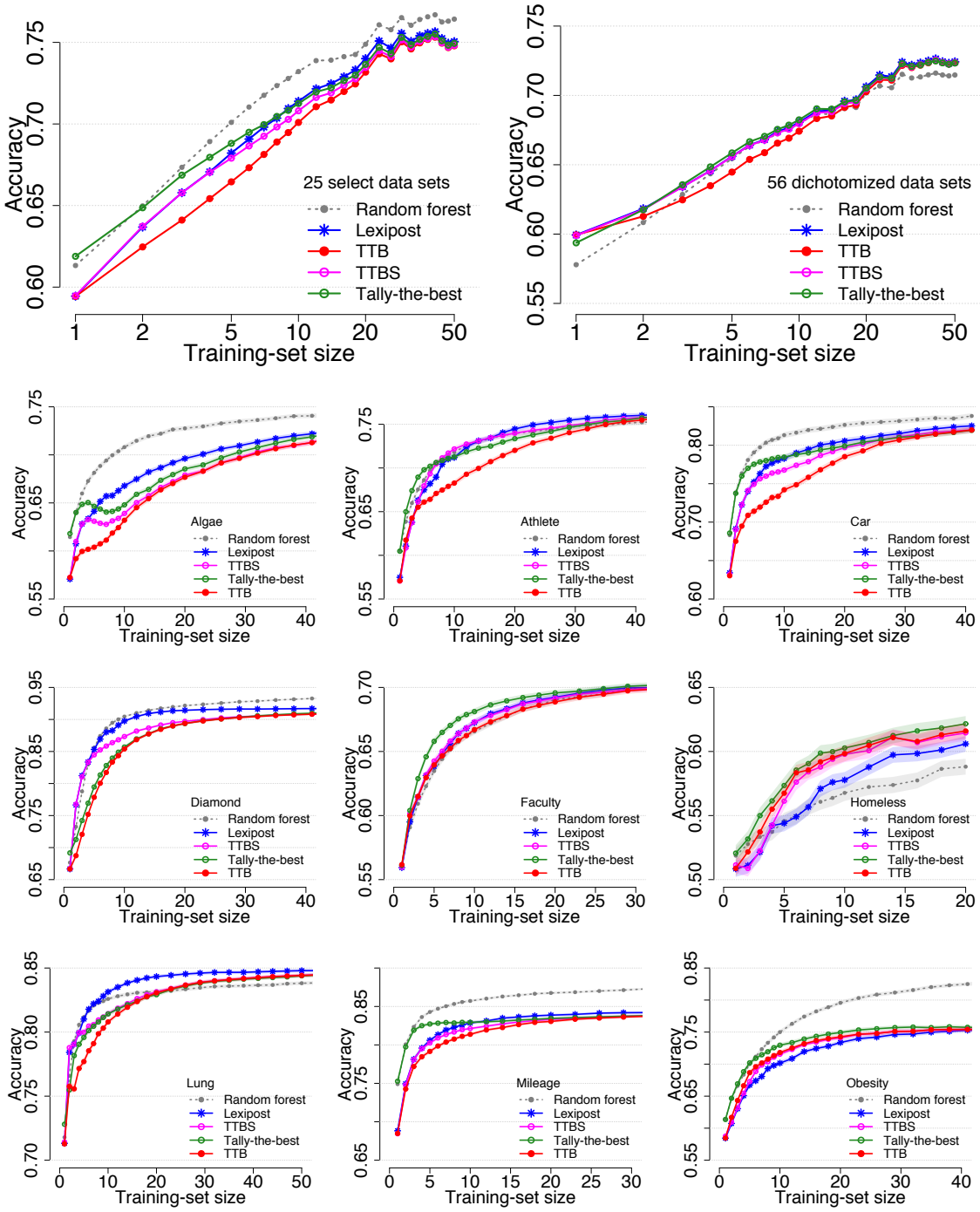

Figure 3: Top left: Mean accuracy in 25 data sets where there was a visible difference between the learning curves of take-the-best (TTB) and lexipost. Top right: Mean accuracy in all 56 data sets, when the cues are dichotomized around the median. In both plots, the horizontal axis is drawn on a log scale. Other plots: Learning curves in 9 of the 25 data sets (no dichotomization of the cues).

7. Discussion

Our results provide a foundation for taking into account sample variance in learning decision heuristics. The far superior performance of lexipost compared to TTB suggests that lexicographic heuristics have large untapped potential. In our simulations, even very simple ways of accounting for sampling uncertainty resulted in large performance improvements in many data sets. Principled methods of handling uncertainty have the potential to further improve performance. Our analysis may be useful in understanding how people take sampling uncertainty into account in learning simple decision rules.

References

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Henry Brighton. Robust Inference with Simple Cognitive Models. In C. Lebiere and R. Wray, editors, *AAAI Spring Symposium: Cognitive Science Principles Meet AI Hard Problems*, pages 17–22, Menlo Park, CA, 2006. American Association for Artificial Intelligence.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G. Goldstein. How good are simple heuristics? In Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group, editors, *Simple heuristics that make us smart*, pages 97–118. Oxford University Press, New York, 1999.

Peter C. Fishburn. Lexicographic orders, utilities and decision rules: A survey. *Management Science*, 20(11):1442–1471, 1974.

Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

Gerd Gigerenzer, Peter M. Todd, ABC Research Group, et al. *Simple heuristics that make us smart*. Oxford University Press, New York, 1999.

Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur, editors. *Heuristics: The Foundations of Adaptive Behavior*. Oxford University Press, New York, 2011.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2. edition, 2009.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002.

Laura Martignon, Konstantinos V. Katsikopoulos, and Jan K. Woike. Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, 52 (6):352–361, 2008.

Özgür Şimşek. Linear decision rule as aspiration for simple decision heuristics. In Chris J. C. Burges, Leon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2904–2912. Curran Associates, Inc., Red Hook, New York, 2013.

Özgür Şimşek and Marcus Buckmann. Learning from small samples: An analysis of simple decision heuristics. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3159–3167. Curran Associates, Inc., Red Hook, New York, 2015.