Estimating Quantities: Comparing Simple Heuristics and Machine Learning Algorithms

Jan K. Woike[1,2], Ulrich Hoffrage[1], and Ralph Hertwig[2]

[1] Faculty of Business and Economics, University of Lausanne, Switzerland
[2] Faculty of Psychology, University of Basel, Switzerland

Abstract. Estimating quantities is an important everyday task. We analyzed the performance of various estimation strategies in ninety-nine real-world environments drawn from various domains. In an extensive simulation study, we compared two classes of strategies: one included machine learning algorithms such as general regression neural networks and classification and regression trees, the other two psychologically plausible and computationally much simpler heuristics (QEst and Zig-QEst). We report the strategies' ability to generalize from training sets to new data and explore the ecological rationality of their use; that is, how well they perform as a function of the statistical structure of the environment. While the machine learning algorithms outperform the heuristics when fitting data, Zig-QEst is competitive when making predictions out-of-sample.

Keywords: estimation, simple heuristics, general regression neural networks, QuickEst, ecological rationality.

1 Introduction

Being able to accurately estimate quantities can be of utmost importance – for instance, for a company that needs to estimate the demand for a new product, or for a government that wants to quantify the effects of legislative changes. An abundance of estimation algorithms have been proposed, some as prescriptive strategies that have been developed to minimize prediction errors, and some as descriptive strategies that have been developed as models of human estimates.

Estimation strategies differ in computational complexity. On the one hand, there are sub-symbolic strategies such as artificial neural nets that require heavy computation. On the other hand, there are much simpler symbolic strategies that could eventually even be executed solely on paper, such as estimation heuristics that have been proposed in the context of the simple heuristics program [1,2]. This program posits that human rationality is bounded, but that cognitive limitations do not necessarily have to be a disadvantage. To the extent that simple heuristics are able to exploit the structure of information in the environment in which humans have to function, they can still reach a high level of performance.

A.E.P. Villa et al. (Eds.): ICANN 2012, Part II, LNCS 7553, pp. 483–490, 2012.

The study of this match between the architecture of decision strategies and environmental structures is central to the study of ecological rationality [3]. In this contribution, two heuristics that have been developed within the simple heuristics framework are pitted against a selection of machine learning algorithms to test for comparative strengths and weaknesses across a range of empirical (data) environments with binary predictor variables (henceforth: cues) and a continuous criterion.

2 Algorithms in the Competition

2.1 Simple Estimation Heuristics

The two heuristics considered here are based on the QuickEst-heuristic, which exploits the fact that in most environments the distribution of criterion values follows a power law [4,5]. The first variant (QEst), requires the following preparatory steps:

1. For each of the k binary cues b_i with possible values 0 and 1, calculate $s_{1,i}$ and $s_{0,i}$ as the average criterion value for all cases o in the learning set (L) with $b_i(o) = 1$ and $b_i(o) = 0$, respectively.
2. Let $s_i^+ = max(s_{1,i}, s_{0,i})$ and $s_i^- = min(s_{1,i}, s_{0,i})$. Recode the cue values such that $s_i^+ = s_{1,i}$.
3. Order the k cues in ascending order of s_i^-, so that for (c_1, ... c_k): $s_1^- \leq s_2^- \leq ... \leq s_k^-$.

QEst can be represented as a minimal binary-tree (see Fig. 1, left side): the tree has k levels following the root node and one exit node on each level following the root node, except for the last level, which has two exit nodes. The cues are assigned in ascending order s_i^- to the k decision nodes beginning with the root node. To create an estimate for a case o, cues are looked up in the order determined above, and once a negative cue value c_j is encountered, an exit node is reached and s_j^- is returned as the estimate. Only if no single negative cue value is found, the mean for all cases in L that do not have a single negative cue value will be predicted.

The only difference between QEst and the original QuickEst is that QuickEst rounds the estimates to the next spontaneous number [4], as it has been designed to model human inferences (and humans tend to generate "round" estimates). Because in an environment in which the criterion distribution follows a power law there are many cases with small criterion values (and these cases will tend to have negative cue values), QEst reduces information search by design and will likely be able to return estimates after inspecting only very few cues. Note that neither of the two variants has free parameters.

The second variant that we introduce and test in this paper, the Zig-QEst-heuristic (ZQ, see Fig. 1, right side), differs from QuickEst on still another dimension: It sorts out the extreme cases on both sides of the distribution first. Cues are put into sequence by choosing the cue with minimum s^- and maximum

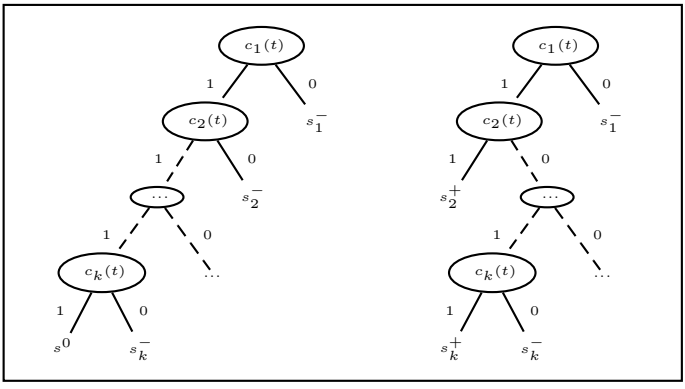

Fig. 1. Structure of the QEst-heuristic (left) and the Zig-QEst-heuristic (right)

s^+, alternatingly (Fig. 1, right side, shows one of the two possible structures). An exit node associated with an s^- is reached when the corresponding cue has a negative value and an exit node associated with s^+, when the corresponding cue has a positive value. The first exit node is placed based on the maximum absolute deviation from the mean $max(|s_i^+ - \bar{y}|, |\bar{y} - s_i^-|)$ across all cues. As a benchmark, the prediction of the mean observed in L was added as a third heuristic. The heuristics were implemented in `Borland Delphi 6.0`.

2.2 Machine Learning Algorithms

The following strategies were chosen from the wide spectrum of algorithms and heuristics that have been proposed in statistics and machine learning as solutions to estimation tasks that involve generalization to unknown quantities. The general regression neural network (GRN) [6] earned the reputation of being able to perform well even for small learning sets (in contrast to, for example, backpropagation networks). This feature is particularly relevant for the present study that uses real-world datasets, which typically do not contain huge numbers of cases and cues. GRN's topology consists of a layer of input units that distribute information to a second radial basis layer, and a final linear layer connected to output units. The radial basis layer consists of pattern nodes that represent the exemplars found in L. Their activation is nonlinearly related to the distance of new cases to these exemplars, and a new prediction is essentially constructed as a blending of values known for exemplars with similar input features. In addition, GRN only has one free parameter, the smoothness parameter σ, which is set to 1.0 in the simulation.

As a second representative of data-mining algorithms, we picked the method of Classification and Regression Trees (CRT) [7]. CRT has been successfully applied to many estimation tasks in pattern classification and estimation. Unlike the first algorithm, CRT does not belong to the class of exemplar-based models. Any instance of a CRT model can be easily translated into rule lists, so that

the application of CRT can be considered as rule-based decision making. Trees are constructed by a partitioning algorithm that recursively separates the set of cases in L into subsets based on single cue values while maximizing their distinctiveness: A common splitting criterion for estimation tasks is the minimization of the predicted squared error, when the mean of cases in the subsets is used as an estimate. Subsets that fall below a minimum size are not split any further. The implemented CRT algorithm has one free parameter: the minimum size for parent nodes that are considered for splits, which is set to 5 in the simulation.

We also included OLS multiple regression (LR) in this category, as it can be considered a standard workhorse in statistical analysis that comes with a proven track record of usefulness in estimation tasks [8]. Finally, as the cues in the task are binary, we added estimation trees (EstT) to the set. EstT make estimations based on a precise match of new cases with known cases: For any profile of cue values that has been observed before, the mean of criterion values for known cases with the exact cue profile is used as an estimate, otherwise the mean of all known cases is predicted. These algorithms were all implemented in `Matlab 7.9`.

3 Simulation Setup

3.1 Environments

The performance of the estimation heuristics and strategies mentioned above has been tested in ninety-nine datasets from various data repositories, on-line data collections, and textbook materials [1,9,10,11,12,13,14,15,16,17,18,19,20]. The datasets were chosen from a diverse range of fields, such as economics, computer science, sports, medicine, social sciences, engineering, and biology. The selection of datasets was completed prior to the first simulation run so that it could not have been influenced by the desirability of results. Continuous predictor variables were binarized by using the mean as threshold. On average, there are 292.2 cases (Range: 11–3450) and 7.4 cues (Range: 2–22) in each dataset.

3.2 Accuracy Criterion

Because the criteria of the ninety-nine environments differed dramatically with respect to their scales, we had to standardize the performance of the algorithms before we could aggregate across environments. Specifically, the performance (A) of a given strategy s in a given data environment d was standardized as

$$A(s) = 1 - \frac{E(s) - E_t}{E_m - E_t}, \tag{1}$$

where $E(s)$ represents the MSE for the algorithm's predictions:

$$E(s) = \frac{1}{m} \sum_{1=1}^{m} (\hat{y}(o_i, s) - y(o_i))^2. \tag{2}$$

E_t is the fitting MSE performance of the estimation tree in the full dataset, that is, with a learning set that consisted of all cases. If $U(x)$ is the subset of cases in L with cue values identical to those of case x, then the prediction of the estimation tree is the mean of criterion values in this subset.

E_t is the lower bound for $E(s)$, if deterministic predictions have to be made for the full dataset based on the cue values, as the mean minimizes MSE for all cue equivalence classes. As the upper bound for the prediction error, in contrast, we take the variance of criterion values in the full dataset ($E_m = \sigma^2(y)$), because this variance is equivalent to the minimal MSE of a model that completely disregards cue information. The criterion $A(s)$ is a linear transformation of $E(s)$ that maps any MSE between these two extremes to the interval $[0; 1]$, such that lower MSEs correspond to higher values on $A(s)$. There may be subsets of the full dataset for which $A(s) > 1$, and ill-fitted models can generate values below 0.

3.3 Learning Conditions

Fitting and prediction accuracy were measured for each strategy in every environment under two conditions: Algorithms were trained and fitted to a learning set of either 50% or 75% of randomly chosen cases from the full dataset and predicted the criterion values in the hold-out set that consisted of the remaining 50% or 25% of the cases in the datasets. All seven algorithms were tested on the same randomly generated subsets, and there were 1000 trials for each combination of dataset, algorithm and learning set size (for a total of $1,386,000$ fitting and $1,386,000$ prediction accuracy results).

4 Simulation Results

4.1 Fitting and Prediction

The results, averaged across the nintey-nine datasets, are presented in Fig. 2, on the left side for fitting, and on the right side for prediction. When fitting known data, the winners of the competition are the machine learning strategies: The best performing strategy is EstT, which is optimally suited to fit data with binary cues (the average accuracy is larger than 1, as it is easier for estimation trees to fit smaller samples). Not far behind in the race are CRT and GRN, followed by LR. In contrast, the two simple heuristics perform much worse, about half way between the machine learning strategies and the mean model as the lower benchmark. As expected, fitting results for the 50% condition are slightly better than for the 75% condition, as it is easier to fit a smaller number of cases with the same number of parameters.

For the prediction task the results are markedly different. First, and as expected, each algorithm performs worse than in fitting. Second, and as expected, results for the 75% condition are better, because larger learning sets yield better parameter estimates. Third, and more interestingly: the estimation tree and CRT are no longer competitive. In the 50% condition, the best performance is

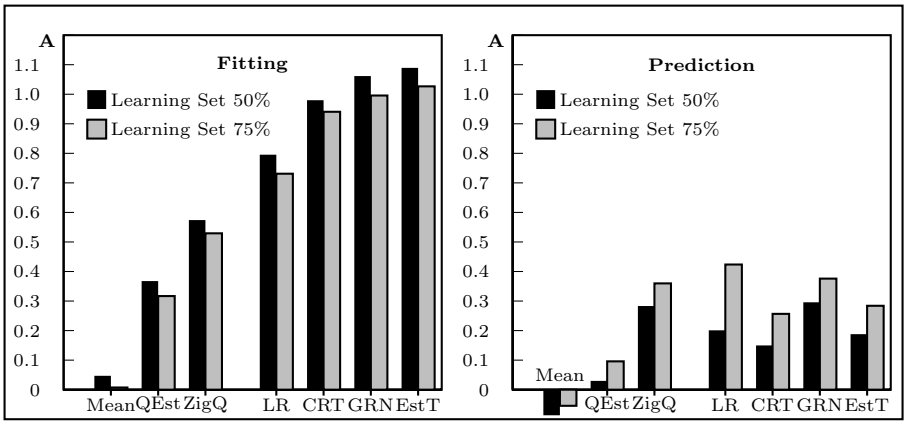

Fig. 2. Average Accuracy-Score (A) for the fitting and prediction of strategies across the ninety-nine data sets: the bar colors represent the learning condition (the size of the learning sets)

reached by GRN and ZQ and in the 75% condition the winner is LR. Both the complex GRN and the simple ZQ-heuristic perform on similar levels, while QEst fails to make good predictions. The difference in accuracy loss between fitting and prediction clearly signals that the complex models over-fit the data. Finally, the difference in performance of LR between the 50% and 75% condition underscores the importance of large sample sizes for generalizable parameter estimates in LR-models.

4.2 Ecological Rationality

The number of datasets used in this study allows for a more detailed analysis of performance differences between strategies across environments, which can be valuable from both a descriptive and prescriptive point of view. In this paper, we will restrict ourselves to report results of a comparison between the most complex algorithm and one of the simple heuristics: GRN and ZQ. Figure 3 shows the difference in predictive accuracy in the 50% condition across datasets, ordered by difference.

Although both algorithms exhibit a similar average prediction accuracy, their relative performance varies across environments. The differences between environments that favor one of the two are not obvious, but some correlation results may shed some light on their nature: the difference in favor of GRN is correlated positively with the number of cases in the datasets (r=.32, p=.001) and negatively with the number of cues (r=-.24, p=.016). The performance of both strategies is positively correlated with the number of cases and negatively correlated with the number of cues. This pattern suggests that GRN can reap larger benefits than ZQ from bigger learning sets and also from a reduction in the number of variables. Further, while the average point-biserial correlation between cues

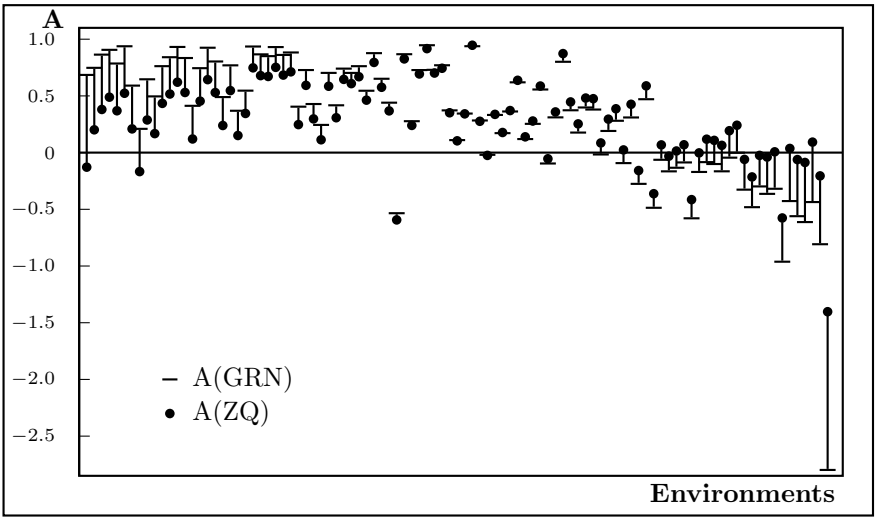

Fig. 3. Predictive Accuracy for GRN and ZQ for each of the ninety-nine data sets ordered by decreasing $\Delta(A) = A(GRN) - A(ZQ)$. The length of the vertical lines corresponds to $|\Delta(A)|$

and criterion in the full dataset is positively correlated with the performance of both algorithms, it shows a positive correlation with the difference between the algorithms (r=.292, p=.003).

5 Discussion

The results of the horse-race simulation clearly demonstrate that predictive accuracy is not necessarily linked to the algorithmic complexity of the strategies. In fact, ZQ, a simple non-compensatory heuristic, which is a plausible candidate for modeling estimation by boundedly rational humans, compared favorably with the machine learning algorithms when the performance across ninety-nine real-world datasets was assessed in cross-validation. The ecological analysis further suggests that the heuristics are less prone to over-fitting, as they suffer less from a decrease in sample size and can cope better with a large number of variables than the machine learning algorithms. The ZQ heuristic drastically outperformed the QEst heuristic and should be more vigorously studied in future research.

These results are in line with previous simulation results for non-compensatory, lexicographic heuristics in pair-comparison [21] and classification [22] tasks. This study also supports the claim that the study of strategies (here, strategies for estimation) cannot be separated from the study of environments in which these strategies are applied [23].

References

1. Gigerenzer, G., Todd, P.M., ABC Research Group: Simple heuristics that make us smart. Oxford UP, New York (1999)
2. Gigerenzer, G., Selten, R. (eds.): The adaptive toolbox. MIT Press, Cambridge (2001)
3. Todd, P.M., Gigerenzer, G.: and the ABC Research Group, Ecological rationality: Intelligence in the world. Oxford UP, New York (2012)
4. Hertwig, R., Hoffrage, U., Martignon, L.: Quick estimation: Letting the environment do some of the work. In: Gigerenzer, G., Todd, P.M., The ABC Research Group (eds.) Simple Heuristics that Make Us Smart, pp. 209–234. Oxford UP, New York (1999)
5. Hertwig, R., Hoffrage, U., Sparr, R.: The QuickEst heuristic: How to benefit from an imbalanced world. In: Todd, P.M., Gigerenzer, G., The ABC Research Group (eds.) Ecological Rationality: Intelligence in the World, pp. 379–406. Oxford UP, New York (2012)
6. Specht, D.E.: A general regression neural network. IEEE Transactions on Neural Networks 2(6), 568–576 (1991)
7. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth, Monterey (1984)
8. Dawes, R.M.: The robust beauty of improper linear models in decision making. American Psychologist 34(7), 571–582 (1979)
9. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
10. Statlib on-line data base, `http://lib.stat.cmu.edu/datasets`
11. DASL - Data and Story Library, `http://lib.stat.cmu.edu/DASL/`
12. OzDASL - Australasian Data and Story Library, `http://www.statsci.org/data/`
13. Journal of Statistics Education Data Archive, `http://www.amstat.org/publications/jse/jse_data_archive.html`
14. Swivel, `http://www.swivel.com/data_sets/`
15. Social Explorer, `http://www.socialexplorer.com/`
16. Inter-University Consortium for Political and Social Research (ICPSR), `http://dx.doi.org/10.3886/ICPSR02650`
17. National Institute for Occupational Safety and Health (NIOSH) Mining Division, `http://www.cdc.gov/niosh/mining/data/`
18. UCLA Statistics Data Sets, `http://www.stat.ucla.edu/data/`
19. Weisberg, S.: Applied linear regression. John Wiley and Sons, New York (1985)
20. Hettich, S., Bay, S.D.: The UCI KDD Archive. University of California, Department of Information and Computer Science, Irvine (1999), `http://kdd.ics.uci.edu`
21. Czerlinski, J., Gigerenzer, G., Goldstein, D.G.: How good are simple heuristics. In: Gigerenzer, G., Todd, P.M., The ABC Reseach Group (eds.) Simple Heuristics that Make Us Smart, pp. 97–118. Oxford UP, New York (1999)
22. Martignon, L., Katsikopoulos, K.V., Woike, J.K.: Categorization with limited resources: A family of simple heuristics. Journal of Mathematical Psychology 52(6), 352–361 (2008)
23. Todd, P.M., Gigerenzer, G.: Environments that make us smart: Ecological rationality. Current Directions in Psychological Science 16(3), 167–171 (2007)