Robust Inference with Simple Cognitive Models

Henry Brighton
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
hbrighton@mpib-berlin.mpg.de

Abstract

Developing theories of how information is processed to yield inductive inferences is a key step in understanding intelligence in humans and machines. Humans, across tasks as diverse as vision and decision making, appear to be extremely adaptive and successful in dealing with uncertainty in the world. Yet even a cursory examination of the books and journals covering machine learning reveals that this branch of AI rarely draws on the cognitive system as a source of insight. In this article I show how fast and frugal heuristics – cognitive process models of inductive inference – frequently outperform a wide selection of standard machine learning algorithms. This finding suggests a cognitive-inspired route toward robust inference in the context of meta-learning.

Introduction

A thorough understanding of how the cognitive system arrives at inductive generalizations would transform machine learning. I will argue that such an understanding would include: (a) a fuller appreciation of the theoretical significance of the multi-task and contextual nature of the inference task; and (b), the realization that cognitive and computational constraints acting on an inference process can act as enablers of robust inference. Within machine learning research, the study of multi-task and meta-learning reflects an important step in the direction of (a). Such studies openly take the cognitive system as inspiration (Thrun & Pratt 1998). In this article I will shed some light on (b) by first showing how a simple cognitive process model frequently outperforms widely used rule-based, connectionist, and exemplar-based learning algorithms. This performance deficiency in existing models is revealed when a cognition-inspired processing limitation is shown to match the structure of the task.

In the context of meta-learning, where base-learners are adaptively selected, diversity in these inference processes is essential. I will argue that cognitive limitations provide a set of constraints that introduce the required diversity in inductive bias, and thereby act as enablers of robust inference. After all, humans represent an existence proof of robust inference, yet are undoubtedly subject to cognitive limitations. The connection between cognitive limitations and inductive bias is potentially significant for both machine learning and psychology. To fully understand this relationship, we require a cognitive-ecological perspective on inference, explaining how properties of the environment can interact with the limitations of the processor. I will argue that the concept of *ecological rationality* provides a route to achieving this perspective, as it focuses on the question of how simple cognitive heuristics can exploit the structure of the environment (Gigerenzer, Todd, & The ABC Research Group 1999). Machine learning and cognitive psychology share a common goal – that of understanding robust learning machinery – but differ drastically in outlook. The work reported here demonstrates and explores a connection: that simple and ecologically rational cognitive process models, built bottom-up, can lead to robust inference.

Background

Within machine learning, inductive tasks are often formalized by first considering an example space $Z = X \times Y$ defining the space of possible observations. An observation is composed of an input $x \in X$ and an output $y \in Y$, where X and Y are termed the input and output spaces, respectively. We also assume some unknown probability distribution $\mu(X, Y)$ on the example space, which will vary from problem to problem. A learning algorithm L takes some finite sequence of n observations, Z^n, and returns a hypothesis H which represents a mapping $H : X \mapsto Y$. Thus, the algorithm L induces a mapping, H, between inputs and outputs over the entire example space Z. A learning algorithm L represents a mapping, $L : \cup_{n \geq 1} Z^n \mapsto \mathcal{H}$, between sequences of examples and hypotheses drawn from a hypothesis space $\mathcal{H}$.

Inductive Performance and Processing Constraints

Satisfactory learning algorithms "should process the finite sample to obtain a hypothesis with good generalization ability under a reasonably large set of circumstances" (Kearns 1999, p159). A useful concept in thinking about generalization ability is *inductive bias*, which is any basis on which one hypothesis is chosen over another beyond mere consistency with the observations (Gordon & Desjardins 1995). Inductive bias can in theory be normative, in the sense that a rational principle such as Occam's razor can be used to choose one hypothesis over another (Hutter 2005). However, achieving adherence to normative intuitions such as these is typically intractable for anything but trivial problems. Restrictions on the hypothesis space arising from issues of computational tractability introduce bias, and when interesting problems meet computationally tractable learning algorithms, this bias tends to be idiosyncratic, reflecting ad hoc properties of the algorithm, and results in the unavoidable conclusion that no single algorithm is the "best" algorithm for a given problem space.

The processes underlying human inductive inference must also operate under certain constraints. But rather than those constraints adopted from the computational complexity hierarchies, cognitive constraints are likely to be quite different and more stringent: such as the physical limitations of the underlying biological machinery, boundary constraints imposed by other cognitive systems internal to the mind, as well as cognitive limitations on, for example, working memory. Faced with the task of making robust inferences, as defined by normative principles of induction, the cognitive system would appear to be facing a tougher problem than that of other forms of computing machinery which are free from idiosyncratic and biologically specific cognitive constraints. But in some respects we observe the opposite. Humans far outstrip machines in their ability, across a diverse set of tasks, to make robust inference about the world.

How to Improve Inductive Performance

Any weak learning algorithm – one which performs marginally better than chance – can be made strong by repeatedly combining the predictions of many such algorithms (Schapire 1990). Here, performance can be improved by using statistical techniques that require costly further processing. For the cognitive system, such a computationally intensive route to achieving robust inference seems unlikely. But how else can performance be improved? Instead of more processing, we can consider less; and instead of generality, we can seek specialization.

The study of fast and frugal cognitive heuristics is perhaps the only mature source of insight into this possibility, where instead of looking to statistically inspired routes to amplify performance, we look to the cognitive system for inspiration (Gigerenzer, Todd, & The ABC Research Group 1999). In the next two sections I will demonstrate how a specialized cognitive model can frequently outperform the standard models of inductive inference. I will then argue that the implication of this result is that meta-learning – the adaptive deployment of inference mechanisms – provides an important connection between machine learning and human learning. In particular, the stock of inductive mechanisms from which a meta-learner adaptively selects can be guided by a cognitive-ecological theory of decision making.

Modeling Inductive Inference

The inductive task to be explored here — the *paired comparison task* — is the problem of choosing which of two alternatives ranks higher on some criterion. For example, choosing which of two houses is likely to be more expensive. The paired comparison task, therefore, is an instance of the widely studied problem of learning from labeled examples.

The two alternatives being considered will be referred to as *objects*. Objects are assumed to have m binary cues, where the cue value 1 indicates the presence of some property represented by the cue, and 0 represents the absence of that property. An *environment* refers to a set of objects and their associated criterion values. For example, a set of houses and their associated prices. In the following experiments, the environment is used to generate these experiences. So rather than representing individual learning experiences, the environment represents an ecological structure which us used to generate the learning experiences. In this sense, the environment has the form a regression problem, but the learning experiences themselves are labeled examples and therefore pose a categorization problem.

To generate experiences from the environment a sample of k objects are drawn at random from the environment. These objects are then used to form pairs of objects, and these pairs represent the specific training experiences. Given these k objects $k(k-1)$ distinct pairs are generated (objects are not compared with themselves). This set of comparisons is termed the training set. In the following experiments, the remaining objects in the environment will often be used to generate a set of experiences used for testing the learner.

One Reason Decision Making with Take the Best

Take the Best is a simple process model of inductive inference for the paired comparison task (Gigerenzer & Goldstein 1996). Take the Best makes an inductive inference on the basis of a measurement taken for each cue, termed the *ecological validity* of the cue. Given m cues, Take the Best computes, for each cue i, its ecological validity, denoted by v_i:

$$v_i = \frac{\text{number of times cue } i \text{ makes a correct inference}}{\text{number of times cue } i \text{ discriminates between objects}}$$

Given a pair of objects, a cue is said to discriminate between the objects if the two objects have different values for this cue. For a cue to discriminate correctly, the object which has the higher criterion value must also have a cue value representing presence of the property represented by the cue. In simple terms, the ecological validity of a cue can be thought of as a measure of how many correct inferences are made (on the comparisons contained in the training set) using this cue alone. On the basis of the cue validities, Take the Best makes an inductive inference between two objects by considering each cue in order of descending cue validity.

When considering the cues in this order, the first cue which discriminates between the two objects, and only this cue, is used to make the prediction; that is, the object possessing the cue indicating the presence of the property is taken to score higher on the criterion. The remaining cues play no part in the decision. For this reason, Take the Best is an instance of a *lexicographic* decision strategy because, rather like looking up a word in a dictionary, a discriminating cue found early in the cue order makes all the subsequent cue values irrelevant. If none of the cues discriminate between the two objects, then Take the Best makes a guess. How can such a strategy, which ignores so much information, compete with sophisticated machine learning algorithms which possess the ability to make inferences on the basis of far more sophisticated relationships between cues?

Five Classic Models of Inductive Inference

To answer this question, I will pit Take the Best against five widely used models of inductive inference. First, I will consider two exemplar models: the basic nearest neighbor classifier (labeled NN1) and a more elaborate model (labeled NN2) based on the GCM model, which uses a weighted function of all stored instances (Nosofsky 1990). Next, I will consider the connectionist approach of using feed-forward neural networks trained using the back-propagation algorithm (labeled BP). Finally, I will consider two classic decision tree induction algorithms: C4.5 and CART. These three families of algorithm span some of the key approaches to modeling inductive inference from both psychological and machine learning perspectives.

Model Comparison

In order to compare the performance of the six models I will use two model selection criteria: (1) cross-validation (CV), to estimate of the predictive accuracy of the models; and (2), the minimum description length principle (MDL), to estimate the degree to which the models compress the training data.

CV. Figure 1(A) depicts predictive accuracy, estimated using CV, of the six models in eight natural environments as function of sample size. Holdout cross validation was used with 5000 random partitions of the environment to generate the training and test sets. These environments are widely available regression problems, and those examined here are a representative sample drawn from 25 test environments examined in a larger study (Brighton 2006). With respect to the performance difference between Take the Best and the five competitors, this sample of eight environments is representative of the performance differences found in the larger study. In Figure 1(A), observe that for half the environments Take the Best clearly outperforms all of the competitors (top row). In the remaining four environments (bottom row) Take the Best performs less impressively, although it often outperforms all the competitors for some subset of the sample sizes. How solid is this finding?

MDL. To firm up the result, I contrasted the ease of comparison but lack of clear justification provided by CV with the more justified but harder to implement criterion of MDL (Pitt, Myung, & Zhang 2002). To ensure a precise comparison, I will narrow down the comparison to include those models which can be interpreted as decision tree induction algorithms. These models are: Take the Best, C4.5, and CART. Figure 1(B), for each environment and model, plots the sample size used to construct the training set against compression rate of the induced hypotheses. The compression rate is the ratio of (a) the length, in bits, of the recoded training data using the induced model, to (b), the length, in bits, of data coded verbatim. The code length functions used are formally identical to those reported in the literature (Wallace & Patrick 1993), and further justified through a derivation based on the modern recasting of MDL (Rissanen 1996; Grünwald 2005). The motivation behind MDL is that the more the induced model helps in compressing the data, the more we have learned about the data. Note that with respect to the model comparison, Figure 1(B) bears an extremely close relationship to Figure 1(A). The two model selection criteria differ drastically in both philosophy and in practice, so the fact that they are in close agreement is very good indication that these results are trustworthy.

This comparison reveals two findings: (1) that Take the Best frequently outperforms some of the most widely studied models of inductive inference within a learning from examples setting; and (2), that Take the Best is doing something quite different and appears to belong to a different breed of mechanism. This second point is illustrated by the fact that the five classic models of inference, in comparison to Take the Best, often achieve very similar degrees of performance.

Discussion

Take the Best has revealed a performance lacuna in some of the key models of inductive inference. In principle, the existence of such a learning algorithm is a theoretic truism (Wolpert 1996). But I will take this performance lacuna to be worthy of further discussion because, significantly, it has been revealed by a cognitively plausible model of inference operating in natural environments. With respect to the five other models in the comparison, Take the Best is relatively task-specific. Its application is restricted to the paired comparison task, and therefore Take the Best makes fairly strict assumptions about Z. Furthermore, and like all learning algorithms, it makes implicit assumptions about target function to be learned, μ.

If inductive performance can be improved by considering simple, cognitively plausible, and specialized processes in this way, then we have a solid basis on which to argue, from both machine learning and psychological standpoints, that collections of appropriately deployed specialized mechanisms can outperform a single "general purpose" mechanism. Indeed, this view is resonant with psychological studies of human decision making where the selection of decision processes, like Take the Best, are seen to vary greatly depending on context-specific properties of the task environment (Payne, Bettman, & Johnson 1993; Bröder & Shiffer 2003, for example). One crucial component assumed here, which I have so far neglected, is that some process governing the selection of specialized processes needs to be specified. Ultimately, such a specification leads to a single meta-algorithm tying together a collection of sub-processes. But, once such a specification is complete, have we then not simply constructed yet another general purpose algorithm, like those explored above?

Base-Learners and Meta-Learners

The models of inference discussed above are *base-learners* and form the basic ingredients for the branch of machine learning concerned with constructing meta-learners (Thrun & Pratt 1998; Vilalta & Drissi 2002). Here, by appropriately

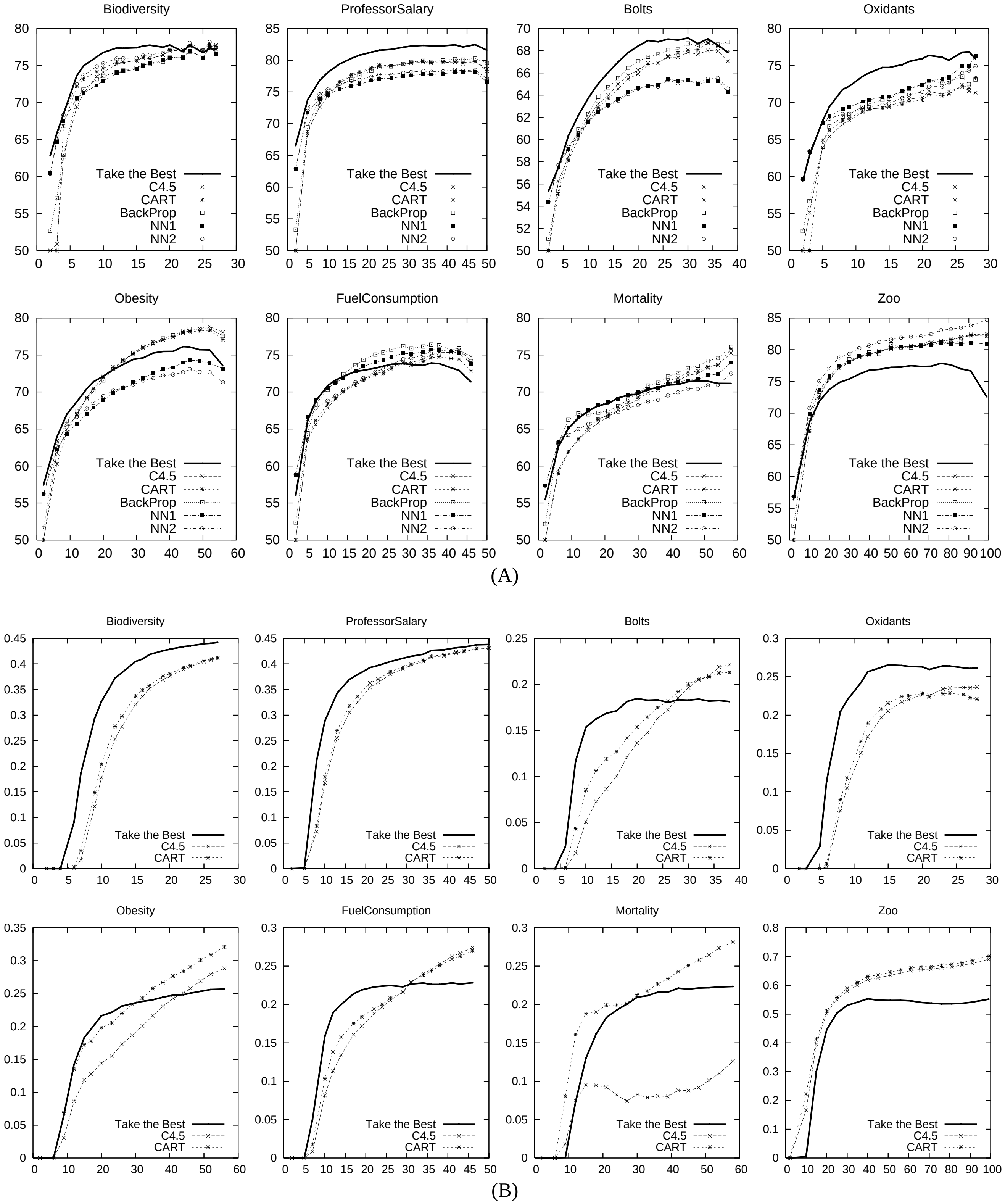

Figure 1: (A) Model comparison using CV. Each plot depicts predictive accuracy as a function of the number of objects used to construct the training set. (B) Model comparison using MDL, where each plot depicts compression rate achieved, using the induced hypothesis, as a function of the number of objects used to construct the training set.

combining or adaptively selecting base-learners with different inductive biases, the performance of the meta-learner can exceed that of any single base-learner. Meta-learning studies typically consider models originally designed as general solutions to learning from labeled examples, and then examines the adaptive deployment or combination of these models.

The Adaptive Toolbox. Similar ideas underpin dynamic models of human cognition, which focus on the selection of cognitive strategies like Take the Best over time (Erev & Barron 2005; Gonzalez, Lerch, & Lebiere 2003; Rieskamp & Otto in press). In this respect, and unlike the machine learning models often recruited as base-learners, Take the Best was never intended as a general purpose model of inference, nor a general solution to the paired comparison task, but rather one process among many residing in the mind's *adaptive toolbox* (Gigerenzer, Todd, & The ABC Research Group 1999; Gigerenzer & Selten 2001).

The metaphor of the adaptive toolbox is sympathetic to the meta-learning perspective: it views the cognitive system, in large part, in terms of a collection of simple heuristics built from cognitive primitives. These heuristics support cognitive tasks such as, for example, making paired comparisons or categorizing objects. Such tasks, according to this view, may be solved by a collection of specialized heuristics each one tuned to particular environment structures. Contrast this outlook with the more typical general-purpose perspective on cognitive processes.

I will take the principle distinction between meta- and base-learners to be the ability of meta-learners to change their internal mechanism contingent on the task. Contrast this ability to that of a base-learner, which will always use the same process (e.g., back-propagation, or decision tree growth) to reach inductive generalizations independent of the specific nature of the task. What theoretical significance could adaptive meta-learning solutions confer? I will consider two consequences of such a design. First, the selective use of inference mechanisms, and therefore the deployment of inductive bias, can lead to improved performance. Second, the use of simple but adaptively selected mechanisms can lessen the processing burden, and thus provide a route to constructing cognitively plausible process models of human inference.

Ecologically Rational Inductive Inference

Machine learning suffers from the lack of cognitive-inspired influences:

> [...] some of the earliest and most influential learning algorithms were developed by psychologists [...] but as machine learning has developed its own identity, the proportion of systems case as serious psychological models has decreased. This trend is unfortunate. Humans constitute our best example of a robust learning system, and using knowledge of their behavior to constrain the design of learning algorithms makes good heuristic sense. (Langley 1996, pp383-384)

I have noted how meta-learning is one area of machine learning which openly takes its inspiration, in part, from the cognitive system and the nature of the tasks faced by it. Humans face a stream of diverse tasks, and this diversity has been proposed as a positive and enabling influence on the ability of humans to achieve robust inference (Thrun & Pratt 1998). But what factors determine the decomposition of the processing problem, and the principles underlying the adaptive selection of these processes?

Satisficing and Bottom-Up Design. Based on Herbert Simon's influential work on bounded rationality, the metaphor of the adaptive toolbox and the associated concept of ecological rationality is one approach to answering these questions. Herbert Simon argued that the cognitive system *satisfices* – finds good enough solutions that approximate rational inference – by matching the structure of the task with the limitations of the processor. Simon then asked the following question: "How simple a set of choice mechanisms can we postulate and still obtain the gross features of observed adaptive choice behavior?" (Simon 1956, p. 129).

The members of the adaptive toolbox, simple heuristics like Take the Best, are candidate responses to this question. Simple heuristics for inference tasks are: (a) models of cognitive mechanisms built from cognitive primitives, or building blocks, rooted in evolved capacities such as recognition memory, cue comparisons, and simple rules to stop search among cues; and (b), designed to exploit the characteristics of natural environments. By making minimal assumptions about the capacities of the cognitive system, heuristics like Take the Best are built through bottom-up design (in constrast to other ecological approaches to understanding the adaptive nature of cognition (Anderson 1990, for example)).

Limitations that Enable. In comparison to the other inference mechanisms compared above, Take the Best is limited. For example, conditional dependencies between cues are ignored. In contrast to the capabilities of decision trees, neural networks, and exemplar models, Take the Best therefore "suffers" from a cognition-inspired limitation. But this limitation proves extremely useful when it meets certain environmental structures. The results of the previous section illustrate this phenomenon. For instance, the other models can occasionally outperform Take the Best, and this demonstrates the advantage conferred by models which act on nested conditional cue dependencies. But in the majority of environments, this extra processing ability leads to poor performance. Take the Best is *ecologically rational* in this context, as its algorithmic simplicity allows it to exploit the environment and outperform those models carrying out more computation.

This demonstration suggests a processing principle which may explain how the cognitive system achieve impressive degrees of robustness in the face of cognitive constraints: The adaptive use of its limitations can be enablers of robust inference (Hertwig & Todd 2003). This is not always the case, as humans are known to err when given inference tasks of particular forms. But if we are interested in engineering learning systems which match the abilities of the cognitive system, then the base-learning mechanisms may profit from being constrained in the same way. This view suggests that the components of a meta-learning systems should not be off-the-peg learning algorithms, but cognitively plausible processes inspired by cognitive limitations. Indeed, to ensure diversity of inductive bias within a meta-learning system, more imagination is needed when designing learning algorithms. Ecologically rational simple heuristics, as I have shown, offer a promising source of diversity.

Conclusion

This observation is inescapable: “almost every problem we look at in AI is NP-complete” (Reddy 1988, p. 15). Consequently, AI is largely the study of “good enough” processing solutions to inherently difficult problems. Inductive inference is no exception. Fortunately, the cognitive system provides a source of insight into how such problems can be dealt with to an impressive degree of success. In this article I have highlighted one way in which machine learning can learn from other branches of cognitive science. Within machine learning, meta-learning is one step in this direction; it is resonant with contemporary theories of human decision making, and promises a route to improved inductive performance.

Effective meta-learning requires base-learners with diverse inductive biases. Rather than reaching into the current stock of learning algorithms, I have argued that cognitively plausible process models of inference tuned to natural environments offer real potential. Importantly, the precise form of the cognitive limitations matter, as they can enable robust inference given the right context. To support this view, I have demonstrated how a simple cognitive process model of inductive inference introduces extra diversity into the pool of potential base-learners: Take the Best frequently outperforms key connectionist, rule-based, and exemplar models of inference.

Acknowledgement. I would like to thank Peter Grünwald for valuable input on the use of MDL with decision trees.

References

Anderson, J. R. 1990. *The adaptive character of thought*. Hillsdale, New Jersey: Lawrence Erlbaum.

Brighton, H. 2006. Ecologically rational inductive inference. Manuscript in preparation.

Bröder, A., and Shiffer, S. 2003. Take the best simultaneous feature matching: Probabalistic inference from memory and effects of representation format. *Journal of Experimental Psychology: General* 132(2):277–293.

Erev, I., and Barron, G. 2005. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review* 112(4):912–931.

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

Gigerenzer, G., and Selten, R. 2001. *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G.; Todd, P. M.; and The ABC Research Group. 1999. *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gonzalez, C.; Lerch, J. F.; and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science* 27(4):591–635.

Gordon, D. F., and Desjardins, M. 1995. Evaluation and selection of biases in machine learning. *Machine Learning* 20:5–22.

Grünwald, P. 2005. Minimum description length tutorial. In Grünwald, P.; Myung, I. J.; and Pitt, M. A., eds., *Advances in minimum description length*. Cambridge: MIT Press. 23–79.

Hertwig, R., and Todd, P. M. 2003. More is not always better: The benefits of cognitive limits. In Hardman, D., and Macchi, L., eds., *Thinking: Psychological perspectives on reasoning, judgement and decision making*. Chichester, UK: Wiley. 213–231.

Hutter, M. 2005. *Universal Artificial Intelligence*. Berlin: Springer-Verlag.

Kearns, M. 1999. Computational learning theory. In Wilson, R. A., and Keil, F. C., eds., *The MIT encyclopedia of the cognitive sciences*. Cambridge, MA: MIT Press. 159–160.

Langley, P. 1996. *Elements of machine learning*. San Francisco, CA: Morgan Kaufmann.

Nosofsky, R. M. 1990. Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34:393–418.

Payne, J. W.; Bettman, J. R.; and Johnson, E. J. 1993. *The adaptive decision maker*. Cambridge: Cambridge University Press.

Pitt, M. A.; Myung, I. J.; and Zhang, S. 2002. Toward a method of selecting among computational models of cognition. *Psychological Review* 109(3):472–491.

Reddy, R. 1988. AAAI presidential address: Foundations and grand challenges of artificial intelligence. *AI Magazine* Winter 1988:9–21.

Rieskamp, J., and Otto, P. E. in press. SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*.

Rissanen, J. 1996. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* IT-42:40–47.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning* 5(2):197–227.

Simon, H. A. 1956. Rational choice and the structure of the environment. *Psychological Review* 63:129–138.

Thrun, S., and Pratt, L., eds. 1998. *Learning to learn*. Boston: Kluwer.

Vilalta, R., and Drissi, Y. 2002. A perspective view and survey of meta-learning. *Artificial Intelligence Review* 18:77–95.

Wallace, C. S., and Patrick, J. D. 1993. Coding decision trees. *Machine Learning* 11:7–22.

Wolpert, D. H. 1996. The lack of a priori distinctions between learning algorithms. *Neural computation* 8:1341–1390.