Relevance and Insight in Experimental Studies

Pat Langley

As its name suggests, artificial intelligence is a science of the artificial (Simon, 1969). As with other conscious creations, there is a great temptation to assume that we can understand the behavior of AI systems entirely through formal analysis. However, the complexity of most AI constructs makes this impractical, forcing us to rely on the same experimental approach that has been so useful in the natural sciences. Many of the same issues and methods apply directly to AI systems, including the need to identify clearly one's dependent and independent variables, the importance of careful experimental design, and the need to average across random variables outside one's control.

However, beyond these obvious features, a compelling experimental study of intelligent behavior must satisfy two additional criteria: it must have *relevance* and it must produce *insight*. We will illustrate these ideas with examples from machine learning, one of the most experimentally oriented subfields within artificial intelligence. Moreover, since AI researchers are often concerned with extending some existing method to improve its behavior, we will focus on this paradigm.

An experimental study of AI methods has relevance if it has implications for problems on which those methods will be used in practice. This criterion is best satisfied by running one's experiments on *natural* domains from the real world. For example, within the machine learning community, most papers report experimental results on data sets from UCI repository, a collection of files that contain data on a variety of natural classification tasks, such as medical diagnosis.

Experiments with natural domains are essential because extensions to existing algorithms, although intuitively plausible, often make little difference in practice. Consider the naive Bayesian classifier, a simple learning method that uses training data to estimate the conditional probabilities of attribute values given the class. Because naive Bayes assumes that each attribute is conditionally independent, given the class, it would seem easy to improve upon by using more sophisticated methods. However, both Kononenko (1991) and Langley (1993) report little or no improvement with extensions to naive Bayes on a number of real-world data sets. Their studies, although giving negative results, were relevant in that they tested their intuitions on natural domains.

However, experimental studies on natural domains alone do not satisfy our insight criterion. The machine learning community, in particular, has come to rely almost exclusively on experiments that compare alternative methods on a variety of standard domains (here 20 or so data sets from the UCI repository), then conclude that one technique or another is superior because it fares better on most of the domains. Such 'bake offs' tell one very little about the reasons for results, and thus do not provide the understanding about causes that we expect in science.

Insight is best obtained by running experiments on synthetic domains that have been designed to test explicit hypotheses, typically motivated by the intuitions behind the original extension. For example, Langley (1993) reports experiments on synthetic domains that involve target concepts with disjoint decision regions, which violate another assumption made by naive Bayes. The importance of using synthetic domains in not because they let one generate some new task, but because they let one vary systematically some dimension of interest, and thus test hypotheses about the conditions

under which one method will fare better than another. Of course, by themselves, studies with synthetic domains do not ensure relevance; Langley found major differences between naive Bayes and his extension on the predicted synthetic domains, but these differences did not carry over to real-world induction tasks.

Thus, truly compelling studies, in machine learning and elsewhere, will include experiments on both natural and synthetic domains, the first to establish relevance and the second to achieve insight. Ideally, they will also *relate* the findings in the two types of study. For instance, if one finds the same shape of results (say when varying some other factor, such as number of training cases) in a synthetic and natural domain, this suggests that the natural domain has similar characteristics to the synthetic one. This strategy lets one move beyond causal accounts in artificial domains toward reasons for success or failure in natural ones, thus giving relevance and understanding at the same time.

Of course, insights about the sources of an algorithm's power are as important as insights about the effects of domain characteristics. Thus, a well-rounded experimental paper will also include lesion studies, which remove algorithm components to determine their contribution, and studies that examine sensitivity to specific parameter settings. Experiments that systematically vary external resources, such as the number of training cases available for learning, should also play a role in any complete empirical study. These recommendations are not new; Kibler and Langley (1988) propose these research strategies in an early paper on the experimental study of learning, and Cohen (1995) makes a broader case for their use with any intelligent system.

We have drawn our examples from machine learning, but the same basic arguments hold across artificial intelligence. Research on planning, natural language, diagnosis, perception, and robotics would all benefit from a more balanced mixture of experiments. Although some work along these lines exists, papers in every AI subfield would be more compelling if they included systematic experiments designed with both relevance and insight in mind. We encourage AI researchers to take both of these criteria into account in their experimental evaluations, and thus to speed progress toward the day when our field becomes a true science of the artificial.

References

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. Cambridge, MA: The MIT Press.

Kibler, D., & Langley, P. (1988). Machine learning as an experimental science. *Proceedings of the Third European Working Session on Learning* (pp. 81–92). Glasgow, Scotland: Pittman. Reprinted in J. W. Shavlik & T. G. Dietterich (Eds.) (1990), *Readings in machine learning*. San Francisco, CA: Morgan Kaufmann.

Kononenko, I. (1991). Semi-naive Bayesian classifier. *Proceedings of the Sixth European Working Session on Learning* (pp. 206–219). Porto, Portugal: Pittman.

Langley, P. (1993). Induction of recursive Bayesian classifiers. *Proceedings of the 1993 European Conference on Machine Learning* (pp. 153–164). Vienna: Springer-Verlag.

Simon, H. A. (1969). *Sciences of the artificial*. Cambridge, MA: The MIT Press.