Machine Learning 3: 5-8, 1988

EDITORIAL

Machine Learning as an Experimental Science

The role of experiments in machine learning

Machine learning is a scientific discipline and, like the fields of AI and computer science, has both theoretical and empirical aspects. Although recent progress has occurred on the theoretical front (see *Machine Learning*, volume 2, number 4), most learning algorithms are too complex for formal analysis. Thus, the field promises to have a significant empirical component for the foreseeable future. And unlike some empirical sciences, machine learning is fortunate enough to have experimental control over a wide range of factors, making it more akin to physics and chemistry than astronomy or sociology.

In any science, the goal of experimentation is to better understand a class of behaviors and the conditions under which they occur. Ideally, this will lead to empirical laws that can aid the process of theory formation. In our field, the central behavior is learning, and the conditions involve the algorithm employed, the domain knowledge, and the environment in which learning occurs. An implemented learning algorithm is necessary but not sufficient: one should also attempt to specify when it operates well and the reasons for that behavior. Lacking theoretical evidence, experimentation is the natural alternative.

As normally defined, an *experiment* involves systematically varying one or more *independent* variables and examining their effect on some *dependent* variables. Thus, a machine learning experiment requires more than a single learning run; it requires a number of runs carried out under different conditions. In each case, one must measure some aspect of the system's behavior for comparison across the different conditions. Below we consider some dependent and independent variables that are relevant to machine learning.

Dependent measures of learning

Most definitions of learning rely on some notion of improved *performance*. Thus, various performance measures are the natural dependent variables for machine learning experiments, just as they are for studies of human learning. Other measures, like 'understandability' of the acquired structures, may also be informative, but these are not relevant unless accompanied by an improvement in performance. In some cases, intuitively plausible learning methods actually lead to *worse* performance (Minton, 1985), so performance measures are central to evaluating almost any learner's behavior.

Many measures of performance are possible. For supervised concept-learning tasks, the most obvious metric is the percentage of correctly classified instances (Quinlan, 1986). One cannot use this metric for unsupervised learning tasks like conceptual clustering, but one can generalize this measure as the average

ability to predict attributes' values (Fisher, 1987). For problem-solving domains, one can examine the number of nodes considered during search (Minton, 1985) or the quality of the resulting solution paths. For grammar-acquisition tasks, one can measure the percentage of correctly parsed sentences and the percentage of correctly rejected non-sentences.

Given a particular performance criterion, one must implement this measure in some fashion. In nonincremental settings, one can present the learning system with a *training* set and then evaluate its performance on a separate *test* set. Preferably, these sets should be disjoint and selected randomly from the available data. For incremental systems, one presents instances one at a time and, after every nth instance, turns learning off and runs the system on a test set. Alternatively, one can treat each instance as both a training and a test datum. In either case, the result is a *learning curve* that shows change in performance as a function of the number of instances encountered. Such curves can be very informative, but one can also condense this information into more succinct summary measures, such as the asymptotic performance and the number of instances required to reach this asymptote.

Varying the learning method

Unlike psychology, machine learning is fortunate in that it can experimentally study the relative effects of 'nature' versus 'nurture.' The simplest way to examine the influence of 'innate' system features on behavior is to compare entirely different learning methods on the same tasks (Schlimmer & Fisher, 1986). Such comparative studies are rare in the literature, but they have an important role to play in our developing science and their frequency should increase with the advent of standard databases.

Even when studying an individual learning method, it is best to place that method's behavior in context. One can usually compare the system's performance to that of a 'straw man' using a simple-minded strategy. In classification domains, one might simply predict the most frequently occurring class; if this covers 90% of the instances, then a learner that achieves 91% accuracy is not impressive. In artificial domains, one can often specify optimal performance as well. For instance, given noisy data with 30% mislabeled instances, a learner that achieves 69% predictive accuracy is actually doing well. Such lower and upper bounds help one calibrate the quality of system behavior.

Given the complexity of most learning methods, finer-grained studies can also examine the effect of specific components. For instance, if a system contains user-specified parameters, one can determine the effect of varying their settings on system behavior (Lebowitz, 1987). Ideally, behavior will be 'acceptable' within a wide range of parameter values, and the same range will work for different domains. Similarly, one can examine the impact of different biases on an inductive learning algorithm or different domain theories on an explanation-based system. Again, negative results can be informative; the system may behave well given any 'reasonable' bias or domain theory.

Some learning systems contain a number of independent operators or components, and one can study each operator's usefulness through 'lesions.'[1] In

[1] This is a common approach in neuroscience, where researchers excise a well-defined area of the brain to determine its role in behavior.

other words, one can run the system with and without a given component, measuring the difference in performance (Schlimmer, 1987). If a component does not aid the overall learning process, then it can be safely omitted.

Although much experimental learning work has focused on inductive methods, one can apply the same methodology to analytic or explanation-based methods. In addition to varying the learning method, one can also control the type and amount of domain knowledge. For instance, more specific domain theories would presumably lead to less transfer and thus slower learning. Future studies should examine the effect of such factors on performance.

Varying the domain characteristics

To study the effect of 'nurture' on a learning system, one must vary the environment or domain in which it learns. Natural domains, such as Stepp's (1984) soybean data, are the most obvious because they show real-world relevance. Also, successful runs on a number of different natural domains provide evidence of generality. However, such environments give little aid in understanding the effects of domain characteristics on learning, since they do not let one independently vary different aspects of the environment. For this, experiments with artificial domains are essential.

For example, noise is an important factor in classification tasks such as learning from examples. Having decided on the 'correct' concept description or decision tree, one can generate instance sets with varying amounts of noise in either the class or attribute information (Quinlan, 1986). Similarly, one can control the complexity of the target concept (e.g., the number of disjuncts) in the given representation language. In the same manner, one can vary the structure inherent in data given to a conceptual clustering system (Fisher, 1987) or the regularity of the problem space given to a heuristics learner. Such domain characteristics may affect learning behavior in significant ways, and undoubtedly other influential features remain to be discovered.

For incremental methods, the order in which one presents instances can be another important factor. Learning curves reflect this influence by treating the number of instances processed as an explicit independent variable. Thus, one way to study order effects is to examine the learning curves that result from different orders. Even when not focusing on such effects, it is important to remember that they may still occur. In these cases, one should collect a sample of randomly selected learning curves and report an average curve.

Designing experiments with learning systems

Basic experimental methodology dictates varying the value of one independent term while holding others constant. However, one can apply this process iteratively to obtain 'factorial' designs in which one observes the dependent measure(s) under all combinations of independent values. This lets one move beyond isolated effects and look for *interactions* between independent variables. For instance, one might find that learning method A behaves better than method B in one environment, whereas B fares better than A in another. Alternatively, one might find interactions between two components of a learning method or two domain characteristics. We believe the most unexpected and interesting empirical results in machine learning will take this form.

In the natural sciences, one can never control all possible variables. As a result, researchers must collect multiple observations for each cell in their experimental design, average the resulting values, and use statistical techniques to ensure that the differences between cells are justified by the data. As a science of the artificial (Simon, 1969), machine learning can usually avoid such complications. Given complete control over the learning algorithm and the environment (if using artificial domains), there is no need for repeated observations or statistical tests. In some cases, as with instance order, practical concerns forbid one from examining all combinations and thus repeated sampling and significance tests are required. However, these are exceptions rather than the rule.

In other words, machine learning occupies a fortunate position that makes systematic experimentation easy and profitable. However, this does not mean empirical researchers should report gratuitous experiments any more than theoreticians should publish vacuous proofs. Whether they lead to positive or negative results, experiments are worthwhile only to the extent that they illuminate the nature of learning mechanisms and the reasons for their success or failure. Although experimental studies are not the only path to understanding, we feel they constitute one of machine learning's brightest hopes for rapid scientific progress, and we encourage other researchers to join in this evolution.

Pat Langley
University of California, Irvine
LANGLEY@CIP.ICS.UCI.EDU

References

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139-172.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning*, *2*, 103-138.

Minton, S. N. (1985). Selectively generalizing plans for problem solving. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 596-599). Los Angeles, CA: Morgan Kaufmann.

Schlimmer, J. C. (1987). Incremental adjustment of representations for learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 79-90). Irvine, CA: Morgan Kaufmann.

Schlimmer, J. C., & Fisher, D. H. (1986). A case study of incremental concept induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 496-501). Philadelphia, PA: Morgan Kaufmann.

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

Stepp, R. E. (1984). *Conjunctive conceptual clustering: A methodology and experimentation.* Doctoral dissertation, Department of Computer Science, University of Illinois, Urbana.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81-106.