Machine Learning 3: 253–259, 1989

EDITORIAL

Toward a Unified Science of Machine Learning

Diversification and unification

Machine learning is a diverse discipline that acts as host to a variety of research goals, learning techniques, and methodological approaches. Researchers are making continual progress on all of these fronts – tackling new problems, formulating innovative solutions to those problems, and devising new ways to evaluate their solutions. Such variety is the sign of a healthy and growing field.

However, diversification also has its dangers. Subdisciplines can emerge that focus on one goal or evaluation scheme to the exclusion of others, and similarities among methods can be obscured by different notations and terminology. Thus, it is equally important to search for basic principles that unify the different paradigms within a field. Just as the twin forces of gravity and pressure hold a star in dynamic equilibrium while generating energy, so the joint processes of diversification and unification can hold a science together while fostering progress.

In this editorial, I examine seven dichotomies that have emerged in recent years to partition the field of machine learning. I begin with three issues related to research goals and evaluation methodologies, then turn to four more substantive issues about learning methods themselves. In each case, I argue that long-term progress will occur only if we can find ways to unify these apparently competing views into a coherent whole.

Accuracy and efficiency

Learning involves some change in performance,[1] and one of the main goals of machine learning is to develop algorithms that improve their performance over time. However, there are many different aspects of performance. For instance, early work on empirical methods emphasized classification accuracy on training sets, while more recent work has focused on transfer of accuracy to separate test sets. In contrast, most work on analytical learning has been concerned with increasing the efficiency of the performance system.

In principle, researchers could continue to pursue these goals independently, but a broader view may lead to deeper insights about the nature of learning. In cognitive psychology, accuracy and efficiency have been two of the main performance measures for decades, and experimental studies have revealed a variety of empirical laws. One of the most interesting relations states that there

[1] Here I am borrowing psychology's distinction between *performance* – an agent's behavior at a given time – and *learning* – the changes in that behavior over time. In this framework, the phrase 'learning performance' is a contradiction in terms.

is a *tradeoff* between the speed at which one executes a skill and the accuracy with which one carries it out. Not all relations take the form of tradeoffs, but they are natural candidates in the search for regularities.

The same notion can be applied to the behavior of intelligent artifacts. Rather than examining performance measures in isolation, one can look instead for relations *between* these measures. For instance, Ellman (1988) has noted a tradeoff between efficiency and accuracy in constructing approximate explanations for analytical learning. Many more such studies will be necessary before we can achieve a unified theory of learning. One need not limit this approach to performance measures; it can be applied equally well to other aspects of learning, such as efficiency of the learning algorithm and complexity of the acquired knowledge structures (Iba, Wogulis, & Langley, 1988; Clark & Niblett, 1989). Researchers should not hesitate to borrow tools and concepts from other fields, such as cognitive psychology and complexity analysis, in pursuing these issues.

Incremental and nonincremental learning

A second goal-related issue involves the distinction between incremental and nonincremental learning. Researchers who explore the former approach are typically concerned with developing plausible models of human learning, with agents that must interact with a dynamic environment, or with the efficiency of the learning mechanism. In contrast, those who employ nonincremental learning methods are typically concerned with automating the process of knowledge acquisition for expert systems.

Despite these differences in motivation, researchers in both paradigms have much to learn from each other. Incremental and nonincremental systems often use the same basic learning operators and produce similar results. In many cases, one can create incremental variants of nonincremental algorithms, as Schlimmer and Fisher (1986) have shown for Quinlan's (1986) ID3 system. Presumably, many incremental learning methods also have nonincremental counterparts.

In this view, the dichotomy is not between different methods, but between different *versions* of the same method, and the interesting questions involve the behavioral differences between these variants. Does the incremental version always acquire the same knowledge structures as the nonincremental one? How many instances does each version require to reach asymptotic behavior? How much total processing time does each use to reach this level of performance? As with accuracy and efficiency, these issues transcend machine learning, and researchers should apply results from complexity analysis and other fields as appropriate. Again, the answers to these questions may involve some form of tradeoff, but the exact relation between such methods is best answered by careful analysis and systematic experimentation.

Theoretical and experimental studies

In recent years, machine learning has made rapid strides along two methodological fronts. New definitions of learnability (Valiant, 1984) and bias (Haussler, 1988) have led to wide-ranging formal results on inductive learning tasks and methods. Over the same period, experimental studies of learning algo-

rithms – on both natural and artificial domains – have led to tentative empirical laws of their behavior (Kibler & Langley, 1988). Both trends are encouraging, since they place the field on a more sound scientific footing.

However, more mature sciences attempt to integrate theory and experiment. For instance, theoretical physicists make predictions that are tested by experimental physicists, and when prediction and observation differ, the theory must be revised. To date, such cooperation between theoretician and experimentalist has been rare in our field, though such exchanges would be a positive development.

Some may argue that machine learning is inherently different from the natural sciences. Because it studies artifacts over which one has complete control, there is no need for experimentation, and formal analysis should suffice. But this view ignores the fact that all theories rely on assumptions that may or may not hold when applied to actual algorithms or real-world domains. Testing one's theoretical predictions through experiments lets one gather evidence in favor of correct assumptions, and it can point toward modifications in the case of faulty ones. Long-term progress in machine learning will depend on such interaction between the theoretical and experimental paradigms.

Justified and unjustified learning

A more substantive issue concerns the nature of the learning process. Empirical learning methods extend a system's original knowledge base, leading it to behave differently on some situations than it did at the outset. Yet such methods involve an inductive leap from instances to general rules or schemas, and this leap is inherently unjustified. No matter how many days the sun rises, there is no proof that it will rise the next day.

In contrast, many analytic methods simply compile the results of a proof into a different form. The resulting rule is justified, in that it does not change the deductive closure of the system's knowledge (Dietterich, 1986). As a result, most analytic techniques have no means for moving beyond the knowledge they are given. The rules they generate may alter their processing efficiency, but these rules do not change the system's external behavior, as do inductive learning methods.

However, both of these criticisms break down on close inspection. At least for humans, very little knowledge of domains is deductively valid; most inference rules are heuristic in nature, with some being more plausible and others less (Collins & Michalski, in press). Because plausibility is not transitive, a compiled rule that is based on many plausible ones may itself be quite implausible. Thus, in general one should empirically test the adequacy of rules learned through analytic methods, and this reduces their distinction from rules acquired through empirical techniques (Pazzani, 1987).

The other claim – that analytic methods cannot lead to behavioral changes – also holds only under unrealistic assumptions. All performance systems have effective limits on their memory and processing time. As a result, the addition of rules that reduce memory load or increase efficiency can allow successful completion of tasks that were not possible before learning (Neves & Anderson, 1981). Thus, analytic methods can lead to changes in external behavior, though in different ways than do empirical techniques.

These arguments do not reduce the very real differences between analytic and empirical learning methods, but they do show that neither approach is superior to the other in any basic sense. Moreover, they suggest that progress lies in the direction of attempts to unify these paradigms within a single framework, rather than in emphasizing one at the expense of the other. There is certainly value in identifying different learning methods and analyzing them in isolation, but such idealizations should be preludes to unified theories rather than ends in themselves.

Knowledge-intensive and knowledge-free learning

A related issue involves the role of knowledge in learning. Some researchers attempt to minimize the knowledge given to their learning systems, aiming for general methods that succeed in many domains. Others follow the expert-systems philosophy, arguing that domain knowledge can (and should) be used to constrain learning, just as it does performance.[2] Clearly, this is more of a continuum than a dichotomy. All learning systems start with *some* knowledge of their domain, even if this consists only of possible attributes and their values. On the other hand, no learning system begins with *all* knowledge of its domain, for there would be nothing left to learn.

This continuum suggests a different set of research questions than the traditional view. Rather than arguing whether the use of background knowledge is desirable, one can examine how learning varies as a *function* of such knowledge. The simplest approach would consider how the *amount* of domain knowledge affects learning rate and other behavioral measures. Theoretical results on inductive bias (Haussler, 1988) and experimental results with empirical methods (Drastal & Raatz, 1988) already exist, but more studies are needed for both empirical and analytical techniques.

Another question involves the *quality* of domain knowledge and its effect on learning. Most work on knowledge-intensive learning has implicitly assumed that the background knowledge is both correct and relevant to the learning task. However, these assumptions will not always hold. We need to develop learning algorithms that can ignore irrelevant knowledge and recover from incorrect biases (e.g., Utgoff, 1986), and we need studies that measure these abilities.

This goal suggests the possibility of learning systems that start with little knowledge and acquire their own domain theories. In fact, many existing methods are *incremental*, and thus alter their knowledge base after each training instance (e.g., Schlimmer & Fisher, 1986). After an incremental induction system has seen a hundred instances, its 'background knowledge' may be quite different than at the outset, and this will affect its response to successive instances. Recent work on representation change (e.g., Schlimmer, 1987) and concept formation (e.g., Gennari, Langley, & Fisher, in press) provide simple examples of this point, but we are still far from methods that can induce the type of domain theories used by analytic techniques. Nevertheless, such approaches begin to blur the distinction between 'knowledge-intensive' and 'knowledge-free' learning.

[2]This second view can be applied both to analytic methods, which transform domain knowledge into some other form, and to empirical methods, which use domain knowledge to rewrite instances in another language.

Cases and abstractions

Yet another dichotomy revolves around the form of knowledge acquired during learning. Traditionally, most researchers have assumed that the learner should store some form of general rules or *abstractions* that summarize experience. More recently, others have proposed the storage of individual instances or *cases*, combined with some form of partial matching scheme that lets one apply these cases to new situations (e.g., Kibler & Aha, 1987). This issue crosses the empirical/analytic boundary, since one can modify either approach to handle case representations.

The case-based approach has highlighted some important concerns that machine learning has historically avoided, including issues of indexing and retrieval. However, the extreme version of this approach also ignores the real advances made over the years in methods for forming and using abstractions. A more interesting approach combines the two paradigms, storing certain cases but also using abstractions to index them and aid in retrieval. Recent work along these lines (Kolodner, 1983; Fisher, 1987) shows that apparently antithetical schemes can be reconciled with little effort, to the benefit of both frameworks. Such integrated approaches provide a good role model for the rest of machine learning.

Symbolic and subsymbolic learning

A final distinction concerns the level at which one represents instances and acquired knowledge. Many researchers in machine learning employ symbolic representations to describe both instances and rules. In some cases, these involve complex logical or relational expressions, but a significant fraction of the work on inductive learning has employed attribute-value or featural representations of knowledge. However, inductive techniques also occupy a central role in other research paradigms, such as neural networks and genetic algorithms. Some researchers in these areas have argued that their methods employ *subsymbolic* representations, which are finer grained and thus more flexible than symbolic schemes (Belew & Forrest, 1988).

However, on closer inspection, the differences between symbolic and subsymbolic systems are more superficial than actual. In many cases, the inputs given to 'subsymbolic' techniques are equivalent to the inputs provided to 'symbolic' induction methods. The input nodes of a neural network correspond directly to the Boolean features often used to describe instances for rule-induction and decision-tree algorithms, and attribute-value representations can be easily translated into the same format. The feature vectors given to classifier systems (Belew & Forrest, 1988) can be mapped across in a similar fashion.

Neither are the hidden units of a neural network any less symbolic than the internal nodes of a decision tree or a concept hierarchy. All can be viewed as functions over the space of instances, and thus represent *concepts*, which are inherently symbolic. The hidden units in a connectionist system need not correspond to English words, but this does not make them any less conceptual or any less symbolic. Similar arguments hold for the bit vectors generated internally by classifier systems. The presence of weights on links or rules does not distinguish these methods either, since many 'symbolic' methods employ them as well (Schlimmer, 1987; Fisher, 1987).

This does not mean that finer-grained representations are useless. Indeed, they may let one acquire concepts unattainable with coarser schemes, and they may lead to powerful emergent effects. However, the real issue is one of the grain size, and not whether the representation is 'symbolic.' Furthermore, there is no special property of 'subsymbolic' methods that make them better able to handle fine-grained representations; traditional 'symbolic' techniques, such as methods for inducing decision trees, can be run on them as well.

There do exist substantial and interesting differences among neural networks, genetic algorithms, and 'symbolic' induction methods. Their learning algorithms, performance elements, and representations of knowledge differ in significant ways, and their inductive biases also appear to be quite different. However, all can be applied to the same class of induction tasks, and they can be compared to one another both experimentally and analytically. Future research on induction should attempt to see beyond the notational and rhetorical differences that divide these paradigms, attempting to understand the relative abilities of each approach rather than claiming at the outset that they are inherently different. Such work may even lead to novel ways of combining the inductive biases of different methods (Utgoff, 1988).

Toward a unified science of machine learning

In summary, research developments in machine learning have led to fuller understanding in many areas, but they have also led to paradigmatic splits within the field. Our emerging discipline has reached a crossroads. We can continue to pursue separate goals, invoke different methodologies, and develop disconnected theories, eventually leading to a wide array of subfields with few common concerns, concepts, or methods. This would not be a terrible fate, since progress would continue, though only in the narrow sense of that term.

Alternatively, we can explore relations between various goals, attempt to combine methodologies, and search for integrated theories of learning that cross the paradigm boundaries which have formed in recent years. Not all learning researchers need devote full time to this endeavor, and in some cases, bridging the gap may involve little more than using new terminology or seeing methods in a new light. Some encouraging signs of cross-paradigmatic research have already started to emerge, but more remains to be done, and I invite experts and novices alike to join in the effort. This quest would result in a broader sort of progress, ultimately leading to a unified science of machine learning.

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

References

Belew, R. K., & Forrest, S. (1988). Learning and programming in classifier systems. *Machine Learning*, *3*, 193–223.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*, 261–283.

Collins, A., & Michalski, R. S. (in press). The logic of plausible reasoning: A core theory. *Cognitive Science*.

Dietterich, T. G. (1986). Learning at the knowledge level. *Machine Learning*, *1*, 287–316.

Drastal, G., & Raatz, S. (1988). *Empirical results on learning in an abstraction space* (Technical Report DCS-TR-248). New Brunswick, NJ: Rutgers University, Department of Computer Science.

Ellman, T. (1988). Approximate theory formation: An explanation-based approach. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 570–574). Minneapolis, MN: Morgan Kaufmann.

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

Gennari, J. H., Langley, P., & Fisher, D. H. (in press). Models of incremental concept formation. *Artificial Intelligence.*

Haussler, D. (1988). Quantifying inductive bias: AI learning algorithms and Valiant's model. *Artificial Intelligence*, *36*, 177–221.

Iba, W., Wogulis, J., & Langley, P. (1988). Trading off simplicity and coverage in incremental concept learning. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 73–79). Ann Arbor, MI: Morgan Kaufmann.

Kibler, D., & Aha, D. W. (1987). Learning representative exemplars of concepts: A case study. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 24–30). Irvine, CA: Morgan Kaufmann.

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: Pitman.

Kolodner, J. L. (1983). Maintaining organization in a dynamic long-term memory. *Cognitive Science*, *7*, 243–280.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

Pazzani, M. (1987). Inducing causal and social theories: A prerequisite for explanation-based learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 230–241). Irvine, CA: Morgan Kaufmann.

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

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.

Utgoff, P. (1986). *Machine learning of inductive bias.* Hingham, MA: Kluwer.

Utgoff, P. (1988). Perceptron trees: A case study in hybrid concept representations. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 601–606). St. Paul, MN: Morgan Kaufmann.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*, 1134–1142.