

The Fifth International Conference on Machine Learning

*Usama M. Fayyad, John E. Laird,
and Keki B. Irani*

Over the last eight years, four workshops on machine learning have been held. Participation in these workshops was by invitation only. In response to the rapid growth in the number of researchers active in machine learning, it was decided that the fifth meeting should be a conference with open attendance and full review for presented papers. Thus, the first open conference on machine learning took place 12 to 14 June 1988 at The University of Michigan at Ann Arbor.

0738-4602/89/\$3.50 © 1989 AAAI.

The conference attracted 320 attendees from over 90 different academic, industrial, and government institutions. Of the 150 papers submitted, 49 were accepted for publication in the conference proceedings (available from Morgan Kaufmann). Of the 49 papers, 20 were presented in three days of plenary sessions during the conference, with the remainder presented at a poster session. Three invited talks were included that reviewed important subfields of machine learning: genetic algorithms, connectionist learning, and formal models of learning. The conference also featured discussion sessions on topics of particular interest to subgroups of the attendees. The discussion topics covered empirical approaches to learning, the sharing of machine-learning data and programs, explanation-based learning, and genetic algorithms. In addition, two receptions were held to provide further opportunity for interaction among conference attendees.

The conference was supported by registration fees and grants from the Office of Naval Research (ONR) Computer Sciences Division, the ONR Cognitive Science Program, and the American Association for Artificial Intelligence.

Papers and Invited Speakers

The 49 accepted papers covered a wide spectrum of machine-learning subfields. The areas included empirical, genetic, connectionist, explanation-based, and case-based learning. Some papers represented hybrid approaches incorporating more than

one type of learning. In addition, papers covered machine discovery, formal models of concept learning, experimental results in machine learning, and the computational impact of learning and forgetting. Empirical techniques for concept learning and explanation-based learning are the two most active disciplines, constituting over 50 percent of the accepted papers.

In this article, we focus our attention on the 20 papers presented at the plenary sessions. In each of the following subsections, we give a brief overview of the content of each of the papers. The scope of the review is limited to reporting the general issues dealt with in each paper as well as any interesting mechanisms employed by the authors in tackling the issues. We do not provide a complete summary of the content of the papers, nor do we evaluate the contribution of these papers to the field.

The conference featured three invited talks. The talks emphasized three subfields of study which have not played a major role in the mainstream of machine-learning research but which are gaining great interest and growing at a healthy rate. These subfields are genetic learning, connectionist learning, and theoretical results in machine learning. See Formal Models of Concept Learning, Genetic Algorithms, and Connectionist Learning for summaries of these talks.

Empirical Approaches to Concept Learning

Programs that induce concepts from examples have become a mainstay of

machine-learning research. Because of its low computational cost and its prior successes, Quinlan's (1986) ID3 program for inducing decision trees is one of the most popular approaches to concept learning from examples. Wirth and Catlett presented a study of the effect of windowing on ID3 performance. Typically, in the presence of large numbers of training examples, one can consider feeding only a reasonably sized data subset to the learner to reduce the algorithm's run time. Given a window size, a decision tree consistent with examples in the window is induced. The resulting tree is then tested on examples outside the window. If the tree fails to classify some examples, the window is expanded (typically to include the exceptions), and the process is iterated. Wirth and Catlett conducted an empirical study of windowing over eight different domains. They illustrated that the iteration over monotonically increasing window sizes resulted in greater computational cost than a single run on all the data. They concluded that in noisy domains, windowing does not appear to have much merit because it incurs greater cost at no advantage in terms of accuracy or correctness.

Utgoff addressed the problem of making ID3 incremental. In domains where examples arrive continually over time, a current decision tree might need to be revised to accommodate future examples. Utgoff presented ID5, an alternative to the incremental ID4 (Schlimmer and Fisher 1986), and the rather simplistic approach that calls for rerunning ID3 over the entire set of examples each time a new example arrives. To avoid losing an entire subtree when ID4 changes the test attribute at its root, ID5 reshapes the tree by "pulling the new test attribute up from below" without discarding any part of the tree. Utgoff ran several tests comparing training cost against the number of examples as well as the learning curves of the different algorithms. He established some improvement over ID4, but improvements over ID3 were not clear. This is mainly due to the fact that the ID3 tree need only be rebuilt when a new example fails to be correctly classified. Thus, not

every new example calls for a revision. Some analysis of algorithm complexity was also provided.

The next two papers in this category presented alternative representation structures for concept learning. Spackman dealt with representing rules in the form of criteria tables. A criteria table, used as a decision rule, consists of a list of a set of n Boolean features. The condition part is of the form: "at least k of the n features are present." In addition to the fact that human experts find such a representation easy to comprehend, it is an effective method for compactly coding a combinatorial number of disjunctive normal forms. Spackman stated that such a representation can only code Boolean functions which possess two properties: unateness and non-equivalence symmetry. Armed with the assumption that such a bias on the representation language is appropriate for his domain, Spackman proceeded to show that his criteria learning system (CRLS) can match the performance of learning systems having a conjunctive bias such as Michalski's AQ15 (Michalski et al. 1986). The major improvement is in terms of the computational cost of running the learning algorithm. Some improvement in accuracy was also attained.

Tan and Eshelman presented a weighted network representation of concepts. The network consists of conjunctive and disjunctive nodes that actually approximate logical conjunction and disjunction using continuous-valued positive and negative activation levels. An initial network with initial weights is created and is then transformed by node merging and weight adjustment. The method is claimed to be appropriate in noisy domains with relatively few examples. The system (IWN) was shown to attain performance levels comparable to those of ID3.

Finally, Cheeseman, Kelly, Self, Stutz, Taylor, and Freeman presented AUTOCLASS. Based on a Bayesian classification model, AUTOCLASS is a conceptual clustering system that does not require the traditional distance measure employed by previous conceptual clustering techniques which appear in the machine-learning literature. Given a data set, a class

model, and the prior probabilities of classes, AUTOCLASS II starts with the assumption that there are more classes than is expected, searches for the best class parameters for that number of classes; and, finally, approximates the relative probability for the number of classes. The number of classes is then decreased and the process iterated until the program converges on a number of classes and the class parameters for which a maximum posterior probability is attained. This maximum is not necessarily the global maximum probability. The result is a probabilistic classification of instances in which class membership is probabilistic rather than categorical (as in more familiar systems in machine learning). The program was applied to several databases with good results. In the domain of infrared astronomy, AUTOCLASS discovered classes that although previously unnoticed by the National Aeronautics and Space Administration analysts, appear to reflect actual physical phenomena in the data.

Explanation-Based Learning

Explanation-based learning (EBL) is a technique for obtaining generalized concept definitions based on an analysis of an example using a domain theory. Not only must the learned concept be general, it must also be operational. Braverman and Russell considered the case where the EBL system has metarules that can access the operability and generality of a concept. These metarules can then be used to control the final concepts that are learned. The most desirable concepts are those at the boundary of operability, that is, the most general operational concepts. In addition to their analysis, Braverman and Russell presented algorithms for finding this boundary under a number of different assumptions.

Rajamoney and DeJong considered problems where multiple, mutually incompatible explanations are possible for a single example. In such cases, the domain theory is an imperfect model of an external domain, so it is not possible to determine the single correct explanation from the domain theory alone. Rajamoney and DeJong

proposed experimenting with the external domain to test out which of the multiple explanations is correct, a technique they call active explanation reduction. They described a domain-independent experiment engine that uses Forbus's (1984) qualitative process theory to represent domain theories. This engine is demonstrated on a problem in chemical decomposition.

Cohen described an approach to the problem of generalizing the number of times an entity is involved in an explanation. For example, a system that was given an example of stacking three blocks in a tower should be able to generalize some aspects of the plan it learns so it can be applied to any number of blocks. Cohen's system, called ADEPT, can be decomposed into a theorem prover, a finite-state machine control module, and an inference algorithm for inducing finite-state machines. By representing the control knowledge learned by EBL as a finite-state machine and then applying the inductive inference algorithm, Cohen's system is able to generalize number as well as find efficient representations of the control knowledge. Although not advertised as such, ADEPT is a hybrid learning system in that it applies inductive techniques to the control representations it learns through EBL.

When EBL is applied to planning problems, the result is often the construction of a macro-operator or plan schemata. Traditionally, the macro-operators are composed of an ordered list of actions. Mooney addressed the problem of learning partial temporal orders for components of the macro-operators, thus providing generality in solving future planning problems. His system has been implemented within the EGGs EBL system and has been applied to a number of domains, including programming, blocks world planning, and narrative understanding.

Hybrid Approaches: Empirical and Explanation-Based Learning

Bergadano and Giordana presented a knowledge-intensive approach to concept induction. Their system, ML-SMART, integrated explanation-based deductive techniques with empirical induction methods. Their method is

essentially explanation based. However, to deal with incomplete theories, which normally result in a failure to generate some explanations, they introduce an inductive step to bridge the deductive gap. The result is an explanation-based system that learns from multiple examples rather than a single instance. The system can thus evaluate the consistency and completeness of the assertions generated during the proof procedure.

Formal Models of Concept Learning

David Haussler of the University of California at Santa Cruz was the invited speaker for this topic. He presented a sampling of the current approaches to formal learning theory. The formal study of learning systems and algorithms is essential to gaining deeper insight into the performance and limitations of learning systems (for example, Haussler's [1987] analysis of LEX). He introduced recent results regarding the size of the training data set for programs that learn from examples. He illustrated the derivation of a formula for the number of examples that are sufficient to probabilistically guarantee desired degrees of correctness and accuracy for a program which learns from examples (see Conclusions for further discussion of this illustration). In addition, he briefly reviewed the Valiant (1984) model of probably approximately correct (PAC) learning (Angluin and Laird 1986); the signal-processing model of learning; and the query model, where a learner can actively seek certain types of examples rather than passively receive them.

Because of its importance to this discussion, we first define PAC learning. A concept-learning algorithm is said to be a $PAC_{\delta\epsilon}$ -learner if, with probability greater than $1-\delta$, the algorithm produces a hypothesis that is no further than ϵ away (according to the distance measure related to probability of error) from the actual concept being learned. Thus, ϵ represents the accuracy, and $1-\delta$ is the reliability of the learner.

Natarajan and Tadepalli provided two frameworks for learning. The first is an extension of the Valiant (1984) framework for learning Boolean func-

tions. In this framework, the learner learns a concept from examples as well as background knowledge (as in EBL [Mitchell, Keller, and Kedar-Cabelli 1986]). One of their results indicates that from the point of view of information complexity, under this framework no extra power is provided by the availability of the background knowledge. The second framework deals with viewing learning as a process by which computational efficiency is improved with experience. This framework contrasts with the concept-learning viewpoint of the first framework. Natarajan and Tadepalli derive conditions sufficient to allow efficient acquisition of heuristics over a restricted class of domains.

Amsterdam introduced two extensions to the Valiant formal model of learnability (Valiant 1984). Some of the restrictions on learnable classes under the Valiant model can be removed by allowing the learner to actively seek examples rather than just passively receive them; in effect, the learner conducts experiments. The other extension involves weakening the Valiant guarantees for concept learnability. This weakening of Valiant requirements suggests the notion of heuristic learnability. In relation to this notion, the density measure for distances between concepts is introduced. The density measure gives a handle on the problem of approximating the target concept with concepts that are "near enough."

Finally, Etzioni addressed the issue of reliable learning. He proposed attaching a hypothesis filter to the output of an arbitrary learner. The filter conducts reliability estimates on the produced hypotheses using a sample population of examples not included in the training set. By deleting the hypotheses that fail to meet the desired reliability and accuracy parameter settings, an arbitrary learner can be transformed into a PAC learner (Angluin and Laird 1986). The results derived are based on theorems in statistics regarding reliable estimation of mean and variance. However, such a scheme can suffer from the problem of deleting too many hypotheses. The second insight of the paper is that PACness can be achieved using statistical testing of hypotheses,

which suggests that the concept-class assumption of the Valiant (1984) model that codes the linguistic bias of the learner might not be a necessary assumption.

Genetic Algorithms

John Holland of The University of Michigan was the invited speaker for this topic. His presentation focused on the genetic algorithm in the context of the classifier system (Holland 1986). He reviewed many successful applications and outlined directions for future research. His outline of future plans included the study of the effects of large systems of classifiers (rule populations of 8000 or more) implemented on the massively parallel Connection Machine. He also mentioned Project 4P, a long-range research program that targets the design of a large-scale cognitive system with a complex environment. The goal is to study the role of genetic algorithms in providing large-scale systems with internal models and lookahead.

Three of the presented papers dealing with genetic algorithms illustrated that much research is needed before the general-purpose search method is clearly understood. Caruana and Schaffer dealt with the interaction between the representation and search mechanism biases. In function optimization tasks, the coding function used to represent the original search function can severely limit the effectiveness of the genetic algorithm search. They illustrated their point by experimenting with six functions for which both binary and Gray codings are used. The results clearly indicate the superiority of Gray coding. They argued that Gray coding improves the genetic algorithm performance by eliminating the Hamming cliffs that make some transitions difficult under the standard binary coding.

Davis and Young addressed difficulties faced by the genetic algorithm because of the exact match procedure used to match classifiers against message strings. They illustrated that for the binary response problem the full power of the genetic algorithm is not brought to bear if the exact match procedure is employed. They proposed

a variation of Booker's (1985) Hamming distance criterion for matching. One extra component, namely, Hamming weights, is added to each classifier. They illustrated that for optimal performance in the binary response problem with level noise, only two classifiers are needed by both Hamming distance criterion and weighted Hamming match. Exact match classifier systems need a combinatorially increasing number of classifiers to achieve a high degree of performance. Weighted Hamming match has the advantage over the Hamming distance criterion when the noise level associated with each bit varies.

Robertson presented experimental results on the use of *CFS, a parallel implementation of a classifier system on the Connection Machine. The primary task domain is letter-sequence prediction. The results indicate that increasing population size (number of classifiers) increases the performance of the classifier system. These results are contradictory to Goldberg's (1985) theory of optimal population size.

Connectionist Learning

Geoffrey Hinton of the University of Toronto was the invited speaker for this topic. He reviewed the different connectionist paradigms including competitive (unsupervised) learning (Grossberg 1987); Boltzmann machine learning; and the recent, familiar backpropagation algorithm (McClelland, Rumelhart, and the PDP Research Group 1986) for training neural networks. He also reviewed new research directions such as *unsupervised backpropagation*, or *generative backpropagation*. The main focus of this research is to overcome the need for the continuous detailed error feedback required in most backpropagation systems. Such a requirement might prove to be a severe liability in domains where complete and continuous feedback is unavailable. According to Hinton, the other important aspect to be addressed in connectionist research is the slow rate at which learning progresses. Even though backpropagation is significantly faster than the "much too slow" Boltzmann machine learning, it is still too slow

for practical applications. He also covered some of the successful implementations of connectionist systems.

Lynne presented a hybrid connectionist learning scheme that combines two ideas from two paradigms of connectionist learning: competitive learning and reinforcement learning. Under his scheme, the competitive learning flavor is retained by using outputs of other units as negative reinforcement (punishment) to a given unit. Reinforcement is also received directly from the environment to supply feedback on the system's performance. Thus, the system can conduct unsupervised competitive learning while benefiting from external advice. Some informal analysis was provided, but it is not clear whether the architecture guarantees stability or correct convergence to the goal.

Machine Discovery

Kelly provided a theoretic framework for studying the effect of the hypothesis language of a learner on the difficulty of the learning problem. Two types of convergence on a true theory were defined: AE and EA.¹ A scheme for classifying hypothesis languages employed by learning programs was established. Several theorems relating the ease or difficulty of learning to the hypothesis language's class were presented. Although some of the assumptions underlying many of the results might not hold in some application domains, the paper did provide a handle on the major role played by the hypothesis language in determining the success or failure of a learning system.

Muggleton and Buntine presented a machine invention system for inventing first-order predicates. The system, CIGOL (LOGIC backwards), invents predicates and conducts generalization on Horn clauses by inverting resolution. This incremental induction is intended to enable the system to formulate its own predicates and augment incomplete clausal theories. Examples of concepts learned include list-reverse, list-minimum, and merge-sort.

Falkenhainer and Rajamoney presented a scheme that integrates a verification-based analogic learning method and an experimentation-based

theory revision method. Based on analogy with prior experience, the former (PHINEAS) forms a theory to explain a phenomenon. The latter (ADEPT) verifies or revises the theory by conducting experiments. The authors demonstrated the synergy and interaction between the two systems as they attempted to explain the phenomena of evaporation and osmosis.

Conclusions

In the following paragraphs, we summarize some of the recurring themes of this conference.

Emergent Themes

Within the papers on empirical approaches to learning, one cannot help but notice the sharing of data sets between researchers. These data sets provide a common yardstick for comparing different approaches. In addition, almost all the new systems introduced were presented in the context of improvement over previous approaches. Comparisons of new and existing systems abounded in the papers accepted for publication. This is a positive feature for an emerging science.

Formal approaches and theoretic analysis of learnability, bounds on performance, and complexity of learning tasks are playing an increasingly important role in giving researchers a greater understanding of the complexity and feasibility of the goals targeted by machine-learning research. Theoretic modeling is becoming a necessity as learning systems increase in complexity and grow in sophistication. An example of the application of theory to practice is what Haussler referred to in his invited talk as Uncle Bernie's rule for concept learning from examples. By fixing a desired error rate, say ϵ , for a concept description to be produced by a learner, it is possible to derive a limit on the probability that the produced concept is bad (that is, has error rate greater than ϵ), as follows:

$$\text{Probability (bad hypothesis is produced)} \leq |H| e^{-\epsilon m}$$

Where $|H|$ denotes the size of the hypothesis space, and m is the number

of examples observed by the learner. Thus, a learner can be probabilistically guaranteed to produce arbitrarily accurate concepts by choosing the appropriate size of the training data set.

Empirical and explanation-based approaches are still the most active subfields of machine learning, with combined (hybrid) approaches beginning to appear.

Connectionist learning and genetic algorithms are regaining interest and appear to be growing at a healthy rate. Planning is under way for large-scale systems in both areas. They are still not well understood approaches, but they appear to be promising directions along the way to building machines that learn.

Perhaps closely tied with the previous point is the emergence of several new representation schemes for learning within a limited variety of domains, thus exploiting special properties of specific tasks. In addition, emphasis is being placed on understanding the effect of the representation scheme on the learner's performance.

Within empirical learning, incremental learning is still being emphasized. Also, attention is being paid to making systems robust in the presence of noise—a direct outcome of attempted applications to real-world domains.

Future Conferences

The machine-learning community has now experimented with both workshops and conferences. With advantages to both, the decision was made to alternate between the two formats every other year. The Sixth International Workshop on Machine Learning will be held at Cornell University from 29 June through 1 July 1989. The workshop will be divided into six disjoint sessions, each focusing on a different theme. The sessions are: Combining Empirical and Explanation-Based Learning; Empirical Learning: Theory and Application; Learning Plan Knowledge; Knowledge-Base Refinement and Theory Revision; Incremental Learning; and Representational Issues in Machine Learning. Each session will be chaired by a different member of the machine-learning community, and will consist of 30

to 50 participants invited on the basis of abstracts submitted to the session chair. Plenary sessions will be held for invited talks.

For more information on the 1989 workshop, contact:

Alberto Segre
Department of Computer Science
Cornell University, Upson Hall
Ithaca, NY 14853-7501, U.S.A.
Email: Segre@gvax.cs.cornell.edu
Telephone: 607-255-9196

References

- Angluin, D., and Laird, P. D. 1986. Identifying k-CNF Formulas from Noisy Examples, Technical Report, YaleU/DCS/TR-478, Yale Univ.
- Booker, L. 1985. Improving the Performance of Genetic Algorithms in Classifier Systems. In Proceedings of the International Conference on Genetic Algorithms and Their Applications, ed. J. J. Grefenstette, 80–92. Hillsdale, N.J.: Lawrence Erlbaum.
- Forbus, K. D. 1984. Qualitative Process Theory. *Artificial Intelligence* 24:85–168.
- Goldberg, D. E. 1985. Optimal Initial Population Size for Binary-Coded Genetic Algorithms, Technical Report, TCGA No. 85001, The Clearinghouse for Genetic Algorithms, Univ. of Alabama.
- Grossberg, S. 1987. Competitive Learning: From Interactive Activation to Adaptive Resonance. *Cognitive Science* 11:23–63.
- Haussler, D. 1987. Learning Conjunctive Concepts in Structural Domains. In Proceedings of the Sixth National Conference on Artificial Intelligence, 466–470. Menlo Park, Calif: American Association for Artificial Intelligence.
- Holland, J. H. 1986. Escaping Brittleness: The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems. In *Machine Learning: An Artificial Intelligence Approach*, vol. 2, eds. R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, 593–624. Los Altos, Calif.: Morgan Kaufmann.
- McClelland, J. L.; Rumelhart, D. E.; and the PDP Research Group. 1986. *Parallel Distributed Processing*. Cambridge, Mass.: MIT Press.
- Michalski, R. S.; Mozetic, I.; Hong, J.; and Lavrac, N. 1986. The Multi-Purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelli-

gence, 1041–1045. Menlo Park, Calif.: American Association for Artificial Intelligence.

Mitchell, T. M., Keller, R. M., and Kedar-Cabelli, S. T. 1986. Explanation-Based Generalization: A Unifying View. In *Machine Learning* vol. 1, 47–80. Boston: Kluwer Academic.

Quinlan, J. R. 1986. Induction of Decision Trees. In *Machine Learning*, vol. 1, 81–106. Boston: Kluwer Academic.

Schlimmer, J. C., and Fisher, D. 1986. A Case Study of Incremental Concept Induction. In Proceedings of the Fifth National Conference on Artificial Intelligence, 496–501. Menlo Park, Calif.: American Association for Artificial Intelligence.

Valiant, L. G. 1984. A Theory of the Learnable. *Communications of the ACM* 27(11): 1134–1142.

Note

1. AE means every expressible question will be answered at some time. EA means there exists a time after which all expressible questions are settled.

Usama M. Fayyad is a graduate student at the Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science at The University of Michigan, Ann Arbor, MI 48109-2110. His Ph.D. work focuses on producing provable improvements to ID3-like decision tree generation algorithms.

John E. Laird is an assistant professor at the Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science at The University of Michigan, Ann Arbor, MI 48109-2110. His primary research interests are in the nature of the cognitive architecture underlying artificial and natural intelligence. His work is centered on the development and use of Soar, a general cognitive architecture.

Keki B. Irani is a professor at the Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science at The University of Michigan, Ann Arbor, MI 48109-2110.