Introduction

This is the fourth and final volume in a series of edited volumes that explore the recent landscape of Learning Systems Research, which spans theory and experiment, symbols and signals. Volume 1 of this series introduced our general focus and goal of exploring the intersection of three historically distinct areas of learning research: Computational Learning Theory (COLT), Neural Networks (NN) and AI's Machine Learning (ML). The second volume focused on specific areas of interaction between theory and experiment. In particular, there is a healthy history of theoretical work that has a continuity with complexity theory in computer science and maintains a life of its own. For many years there has also been a very empirical enterprise known as Machine Learning. Until recently, these two trends had generally diverged and done little to inform one another. Volume 2 was an attempt to provide clear examples of how theory can be tested with the careful experiment and how empirical work can be guided in meaningful ways by theory. Recently, there has been more work towards testing theory and rationalizing experiments. We hope these volumes contributed to that trend, at least in some small way.

The final two volumes (including the present one) focus on certain key areas of learning systems research that have developed recently. Volume 3 provided examples of how researchers have conceptualized the problem of "Selecting Good Models." This is a central problem, as any learning approach must specify, identify or select a model from a set of possible models, given data arising from the world and some prior beliefs about the how the data may have been generated.

This volume explores the terrain of "Making Learning Systems Practical." Although there are many factors that may render a learning system useless or impractical, the submissions to this workshop identified three general problems as critical: (1) Scaling up from small problems to realistic ones with large input dimension (e.g., > 10) and examples (e.g., 1000s); (2) increasing efficiency and robustness of learning methods; and (3) strategies to take in obtaining good generalization from limited or small data samples. These, plus a section on specific examples, form the four categories in this volume.

The earlier history of Machine Learning Research was motivated by problems that focused on the basic complexity of learning concept descriptions, rather than on their scaling or potential statistical properties. For example, Winston (1975) illustrated that it was easy to identify the target concept if the training set included (carefully chosen) examples that were just at the decision boundary between the positive and negative sets. This telling example set the stage for both interest in pedagogy (i.e., intelligent tutors) and learning theory. Others (Mitchell, 1980; Michalski & Stepp, 1983) discussed the difficulty of learning complex concepts without significant bias in the representation language. These trends tended to supplant more experimental approaches of using difficult real-world problems (e.g., robotic control, optical character recognition, speech recognition) until the mid-1980s when Neural Networks exploded into the Machine Learning scene (Rumelhart et al., 1986). Unlike Machine Learning, which tended to define its own problems, the field of Neural Networks (in order to transcend its checkered past) had to solve, or at least do as well as any methods on, some hard problem, such as speech recognition; here, this often means competing against approaches that had been fine-honed for 20 years (e.g., HMMs). This orientation toward difficult problems has set standards that any learning method now has to meet in order to be taken seriously.

A current focus in learning research is identifying what makes a learning task difficult and then tracking how the performance of the learning system varies as this aspect increases. This "scaling problem" is the main focus of a number of papers in this volume: **Banerjee** (chapter 1) shows that initializing a neural net using a decision tree can improve both the system's accuracy and efficiency. In a similar vein, **Obradovic** (chapter 2) evolves his system online, by increasing the number of elements in the network's hidden layer whenever the current topology cannot capture the dynamics of the new data. **Shklar & Hirsh** (chapter 3) show how to increase the efficiency of a well-known conceptual clustering approach by supplying information about the number of expected clusters in the domain. Finally, **Langley & Sage** (chapter 4) provide a general analysis of learning rate of two algorithms in the presence of increasing number of irrelevant features.

Other articles directly attack some problem that makes scaling hard — in particular, the robustness and efficiency of the learning system. Again, few if any of the early Machine Learning researchers considered the robustness of their learning system against, for example, irrelevant or noisy features. The focus, instead, was on the nature of the algorithm's behavior on data obtained in the best of worlds; hence these early machine learning systems were almost always given complete, consistent and noise-free data sets — i.e., every observation is composed of all-and-only the relevant features — all feature values are included, and each value is 100% correct. Of course, data is rarely so complete in the real world. Similarly, many machine learning researchers did not appreciate that their training data might be only a sample from a larger population, or even from a potentially infinite data generation sources; nor that the asymptotic analysis from statistics always favored acquiring more data in order to achieve the most robust estimates of the learning systems parameters. However, it is also rare for a real-world learning system to have access to a sufficient amount of data to identify uniquely the target concept that was the source of the data.

Efficient use of data invokes a tradeoff between exact fitting of the limited sample now present, versus extrapolation to any out-of-sample data not yet seen. If the sample is too small, sampling errors will bias the learning system to be overly sensitive to the noise present and therefore produce errors on other unseen samples. On the other hand, if the learning system underfits the training sample, it runs the risk of missing relevant patterns in the data that suggest the nature of the process that produced that data. Maximum Likelihood is one approach to resolve this tradeoff: here, the data is conditioned on the model which in turn is maximized in order to produce parameter estimates of the model that are most likely to have given rise to the observed data. Assuming a parameterization of the model contains the target concept, Maximum Likelihood can be a very powerful approach.

A number of papers in this volume deal with robustness and efficiency. Two papers deal explicitly with learning classifiers that can handle missing attribute values: **Ghahramani & Jordan** (chapter 5) discuss how the EM algorithm can cope with such omissions, and **Schuurmans & Greiner** (chapter 6) formally investigate the various ways in which attribute values can be blocked to determine when (PAC) learning is possible. **Towell** (chapter 7) considers learning from training data that includes both labeled and unlabeled examples. **Liu** (chapter 8) considers robustification approaches that allow learning from a sample that includes inappropriate data. The final two papers in this set deal with data efficiency: **Deco & Schürmann** (chapter 9) show how to dealing efficiently with given data in a difficult chaotic time series; and **Schuurmans & Greiner** (chapter 10) define "sequential learning," a new paradigm for learning, and show it can be much more sample efficient than standard PAC results.

In some sense scaling, robustness and efficiency all depend on our understanding of a learning system's ability to generalize correctly to unseen examples. The analysis and understanding of generalization then poses a key problem for how learning systems could be made practical. Unless we can specify how well our learning system will generalize to out-of-sample cases, there is little chance the system will be accepted in standard engineering domains, where the predictability of the system is crucial.

Once again, we can compare the present focus to early concepts in the machine learning field. Initially, the notion of generalization was neither appreciated nor studied. The main focus of the early Machine Learning researcher was to provide algorithms that could correctly induce the underlying concept that was generating the examples shown it. Whether the system would generalize was irrelevant at that point, since once it had correctly acquired the underlying concept, it could then use this concept to label any further examples. This research bias was most likely inherited from language learning research, where theoreticians had decided that the benchmark problem was building learners that, when exposed to sentences produced by a grammar, could correctly identify the grammar. Here, the learner had to be errorless. Further analysis of such cases showed that, unfortunately, learning grammar with no errors and with perfect probability was usually intractable. This was the dominant model until the mid-1980s, when Valiant introduced in the so-called "PAC learning" paradigm (Valiant, 1984). This model weakened many of the language learning assumptions, allowing the learner to make some errors, provided the errors were usually very small. As learning under this model is tractable in many domains (such as language learning), many machine learning practitioners are beginning to investigate and appreciate other statistical interpretations of learning.

Presently, generalization within a learning system is an intense area of study from both theoretical points of view (see the work on Vapnik-Chervonenkis dimension (Vapnik & Chervonenkis, 1971b), Bayesian learning (Haussler et al., 1992; Neal, 1993; Cheeseman et al., 1988), Statistical Mechanics (Seung et al., 1992), etc.) and from many empirical settings (cross validation (Moore & Lee, 1994; Craven & Wahba, 1979; Liu, 1993), regularization (Shizawa, 1993; Moody, 1992), etc.). In this volume, generalization is studied in a number of model systems. **Murphy & Pazzani** (chapter 11) investigate whether the smallest decision tree necessarily has the smallest generalization error; **Rao & Oblow** (chapter 12) consider methods for fusing the results of N learners, showing when this can approach the optimal Bayesian fuser; **Zador & Pearlmutter** (chapter 13) study the computational capacity of dynamic systems, especially systems defined by networks of spiking neurons; **Golea** (chapter 14) presents an average-case analysis of unions of non-overlapping perceptrons; and **Garvin & Rayner** (chapter 15) probabilistically analyzes combinations of Volterra expansion functions. Finally, **Klenin** (chapter 16) provides a new set of tools for the analysis of generalization, especially related to the variance in generalization error, which can be calculated directly in many systems.

The last, and definitely the most important, section of the book presents real-world learning systems themselves: **Bhansali & Harandi** (chapter 17) show how to use derivational analogies to speedup a system that synthesizes UNIX shell scripts from program specifications; **Almuallim, Akiba, Yamazaki & Kaneda** (chapter 18); show how to learn rules for translating Japanese verbs to appropriate English verbs; **Brāzma** (chapter 19) presents a technique for restoring a regular expression from a given "nearby" string of characters and shows how this can apply in analysis of biosequential data; **Schlang, Abraham-Fuchs, Neuneier & Uebler** (chapter 20) investigates ways to classify biomagnetic fields; and **Vaina, Sundareswaran & Harris** (chapter 21) present a computationally realistic model of human motion detection.

This volume completes the four-volume "Computational Learning Theory and 'Natural' Learning Systems" series, which has covered the vast array of theoretical and empirical results across the multiple fields represented by Machine Learning, Neural Networks and Computational Learning Theory. It thus contains a fairly comprehensive snapshot of the mainstream learning research over the last 5 years and, to that end, encompasses the trends (and perhaps suggests the future) of the dynamic landscapes of Learning System Science.