S. J. Hanson W. Remmele R. L. Rivest (Eds.)

Machine Learning: From Theory to Applications

Cooperative Research at Siemens and MIT

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Series Editors

Gerhard Goos
Universität Karlsruhe
Postfach 69 80
Vincenz-Priessnitz-Straße 1
W-7500 Karlsruhe, FRG

Juris Hartmanis
Cornell University
Department of Computer Science
4130 Upson Hall
Ithaca, NY 14853, USA

Volume Editors

Stephen José Hanson
Siemens Research Center, 755 College Road East
Princeton, New Jersey 08540, USA

Werner Remmele
Siemens AG, ZFE BT SE 2, Otto-Hahn-Ring 6
W-8000 München 83, Germany

Ronald L. Rivest
Room 324, MIT Laboratory for Computer Science, 545 Technology Square
Cambridge, Massachusetts 02139, USA

CR Subject Classification (1991): I.2, F.1.1, C.1.3

ISBN 3-540-56483-7 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-56483-7 Springer-Verlag New York Berlin Heidelberg

Printed in Germany

Typesetting: Camera ready by author/editor
45/3140-543210 - Printed on acid-free paper

Preface

This volume includes some of the key research papers in the area of machine learning produced at MIT and Siemens during a three-year joint research effort. It includes papers on many different styles of machine learning, spanning the field from theory, through symbolic learning, to neural networks and applications.

The joint research cooperation between MIT and Siemens began in 1987, when both organizations independently became interested in pursuing research in machine learning, because of recent technical advances and perceived opportunities. Siemens desired to establish a world-class Learning Systems Laboratory for the long-term purpose of developing corporate expertise and building applications in this area. MIT's Laboratory for Computer Science desired to strengthen its research focus in learning through new research in the area, as well as increased coordination with related research in MIT Artificial Intelligence Laboratory and other parts of MIT. In addition, Siemens looked to MIT for scientific leadership, while MIT looked to Siemens for its potential for industrial applications. We believe that our joint efforts have contributed substantially to advancing the state of the art.

The joint research program has exhibited a diversity of objectives and approaches, including, among others, natural and artificial connectionist learning methods, learning by analogy within a knowledge-based system, learning by the simulation of evolution, the theoretical study of concept learning from examples, and the learning of natural language.

During the first three years of the joint effort, some 60 papers were published, three workshops and conferences were held, and many visits and exchanges of personnel took place. This book contains a sampling of the research produced.

Siemens AG, Munich
MIT, Cambridge, MA
December 1992

Professor Dr. Heinz Schwärtzel
Professor Michael Dertouzos

Table of Contents

Strategic Directions in Machine Learning

Stephen José Hanson[1], *Werner Remmele*[2*] *and Ronald L. Rivest*[3†]

[1] Learning Systems Lab.
Siemens
755 College Road East
Princeton, NJ 08540

[2] Siemens AG
Corporate Research and Technology
Otto-Hahn-Ring 6
D-8000 Munich 83
Germany

[3] MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139
USA

Learning Systems research has been growing rapidly in three historically distinct areas: *computational learning theory*, which has undergone a renaissance in the last few years, *connectionist/neural network learning*, which has seen exponential growth, and *symbolic machine learning*, which has become a dominant influence in the field of AI. All three areas maintain their separate conferences and activities, but they overlap in significant ways at the conceptual level, supporting the hope that these areas will continue to interact and cross-fertilize each other.

Potential barriers to congenial commerce between these areas can be understood better by considering two dimensions of potential interaction. One relevant dimension is the *experimental versus theoretical* approach to learning systems. Theory dominates one of the three areas but it appears sparingly (usually independently of experiment) in the other areas. A second dimension is the *signals versus symbols* axis: continuous variables versus discrete variables, dynamical systems versus logic, or numbers versus words. This second dimension represents a comparatively older tension arising between traditional computer science and engineering (including AI) with classical statistics, control theory, and neural networks.

Illustrating the first dimension (theory versus experiment), computational learning theory research has had little or no contact with experiment; such research tends to focus on the general computational or sample size requirements of a particular learning problem or learning algorithm. This general focus tends to limit the contact of theoreticians with experimental work, which usually focuses on achieving the best possible results in a specific case. In contrast, both

* W. Remmele's network address: remmele@ztivax.zfe.siemens.de
† R. Rivest's network address: rivest@theory.lcs.mit.edu.

neural network learning and symbolic learning have been almost exclusively empirical fields. Moreover, computational learning theory has provided a surprising number of negative results, illustrating that many learning problems are indeed difficult in their full generality, while both of the more experimental learning areas have many positive (but modest) results for learning in particular problem domains. This apparent discrepancy has caused many experimentalists to ignore theoretical developments.

On the other hand the experimentalists have tried many kinds of algorithms and representations that, to a theoretician, are blatantly ad hoc. The experimentalist's choice of a particular set of discrete or continuous features defining fuzzy, probabilistic categories can seem arbitrary to a theorist concerned about more generic learning problems. To the experimentalist such differences loom large; in one case the learning may concern human faces and in another case concern the verb argument structure of natural language, wherease both cases may to a theorist reduce to "concept learning." These differences in focus create guls between experimentalists and theorists that can lead to miscommunication and confusion.

Along the second dimension in recent years we have also seen conflicts between symbolic and connectionist approaches to learning. These approaches could perhaps be considered as based on different computational styles, leading to different accounts of similar phenomena. At one level the conflicts are quite real and yet at slightly more abstract views we see many commonalities. Symbolic approaches have stressed heuristic, deterministic, and deductive models while connectionist approaches have stressed optimal (sometimes heuristic) stochastic models. Such differences do not necessarily strictly polarize researchers but can lead to significant disagreements on the kind of approach that is most likely to work for a given application or problem. Fortunately, in experimental domains differing approaches can be tried on the same data sets, yielding insights about the relative strengths of the approaches.

The present volume attempts to examine the state of the learning field by looking at representative research examples from each of the three areas. These examples should provide the reader with a rich view of the three fields and their potential for integration. There are several themes that emerge from the joint consideration of these different learning approaches; we feel these themes are somewhat generic and provide future directions for research in learning systems. The following is a short list of such themes and questions; many of them arise within the context of the present volume.

What are natural systems? Alternatively how can we make learning easy? Does the world conspire to make learning easy for us by providing "good" examples, examples at the right time, or sets of examples that would be unlikely for other kinds of tasks? Are biological systems wired up in a certain way to advantage of the way the world is structured? Are there simple constraints on classes of functions that can generically improve their inductive bias? Are neural net function classes an example of this improved bias? How can we characterize these function classes and biases in order to construct learning systems that are likely to learn as easily as and with the generality of biological systems?

How should learning systems gain from prior knowledge? Many researchers agree that the trick is to walk a reasonable line between "giving away the store" by building in an unreasonable amount of prior knowledge, and forcing our systems to restart every new learning task from scratch. For example, some researchers have shown that neural networks can converge more rapidly when prior knowledge can be translated into an initial set of weights. These researchers have also tried to estimate the number of examples equivalent to the given prior knowledge.

What makes a learning problem hard? In trying to characterize the kinds of learning tasks that are well suited to networks or symbolic algorithms, one must account for the fact that some algorithms seem to work better with one kind of data than with another. Some learning researchers have argued that the style of computation is critical in this regard. Some problems are just better solved with a neural net while others seem more suited to rule-based symbolic method. What dimensions characterize these biases? This question is rather difficult to be answered; even after years of research, no sufficient set of criteria is yet known that characterizes these biases.

Theoretical studies also highlight the importance of *computational complexity* and *sample complexity* (reflecting the initial uncertainty of the learner) in understanding the "hardness" of particular learning problems.

If knowledge is important, can we quantify how important? We know from statistics how confidence in a hypothesis varies with the amount of data used to support it. Or: if picture is worth a 1000 words, how much prior knowledge is worth 1000 examples? An open question is how the confidence in a hypothesis should vary with both the prior knowledge and the given data. Or: how convincing is the data really, when the hypothesis is seen in light of what we already knew?

How are neural network learning and symbolic machine learning similar? Methods in symbolic machine learning such as constructive induction that create new features can be likened to neural net methods that develop higher-order features in the hidden layers of a neural network during training. Constructive induction methods require either some prior knowledge of potentially useful features or ways to build them, which can be a liability if the domain is truly knowledge-free, but an advantage when we know a little, because they permit a direct encoding of that knowledge. Many researchers agree that there are important similarities between neural nets and symbolic machine learning methods that need further exploration.

How can we trade off complexity of hypothesis with fit to data (Occam's razor)? This is similar in statistics to the notion of trading estimation error for approximation error. Without removing noise or systematic errors, our parameter estimation techniques tend to bias away from the true underlying data model while those same techniques given enough resources (training examples) will approximate the data perfectly. Thus, we must strike a balance between accounting for a known data sample and pursuing constraints (e.g., prior knowledge) on the approximate model that better represents the true model. This tradeoff is also illustrated by learning systems that try to adjust the complexity of their representations while learning from examples.

We believe these questions can help frame the intersection of these three areas and drive them towards some common set methods and views. Further, we feel that, as illustrated by the research presented in this book, the present outlook for machine learning is very favorable and exciting. As machine learning research progresses we expect an increased overlap and synergy among the three fields, leading to learning methods that are both founded in a secure theoretical understanding and successful in practice.

Part I

Theory

Introduction

Ronald L. Rivest

MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139
USA
rivest@theory.lcs.mit.edu

This part of the book includes three papers on theoretical aspects of machine learning. The first two papers use computational complexity-theoretic techniques to derive some fundamental limits on what is efficiently learnable. The third paper provides, by contrast, a positive result—an efficient algorithm for identifying finite automata.

Avrim Blum and Ronald L. Rivest, in *Training a 3-Node Neural Network is NP-Complete*, show that training even very simple neural networks may necessarily be computationally very expensive. The result suggests that researchers may have to be satisfied with the performance of approximate neural-net training heuristics (like back-propagation) or that they should search for alternate representations that are easier to train.

Michael Kearns and Leslie Valiant, in *Cryptographic Limitations on Learning Boolean Formulae and Finite Automata*, show that learning certain concepts classes is no easier than breaking certain well-known cryptographic systems. This provides evidence that these concept classes are intrinsically difficult to learn. Interestingly, and unlike the results of the previous paper, these results are *representation-independent* in that they remain valid no matter how the concepts being learned are represented.

Ronald L. Rivest and Rob Schapire, in *Inference of Finite Automata Using Homing Sequences*, show that *active experimentation* can provide an effective tool for learning finite automata, even if the automaton has no "reset" capability. The ability to perform experimentation provides a way around the limitations proved by Kearns and Valiant in the previous paper. Extending prior work of Dana Angluin, the authors provide here a polynomial-time algorithm for identifying an unknown finite automaton using experiments and "equivalence queries."

These theoretical results illustrate the key role theoretical studies can play in clarifying the effects of changing the underlying model of learning, and in helping to distinguish, in a precise sense, what is efficiently learnable from what is not.