Lecture Notes in Artificial Intelligence 904

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Paul Vitányi (Ed.)

Computational Learning Theory

Second European Conference, EuroCOLT '95
Barcelona, Spain, March 13-15, 1995
Proceedings

Springer

Series Editors

Jaime G. Carbonell
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891, USA

Jörg Siekmann
University of Saarland
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

Volume Editor

Paul Vitányi
CWI, Kruislaan 413
1098 SJ Amsterdam, The Netherlands

CR Subject Classification (1991): I.2.6, I.2.3, F.4.1

ISBN 3-540-59119-2 Springer-Verlag Berlin Heidelberg New York

CIP data applied for

Typesetting: Camera ready by author
SPIN: 10485503 45/3140-543210 - Printed on acid-free paper

Preface

The SECOND EUROPEAN CONFERENCE ON COMPUTATIONAL LEARNING THEORY (EuroCOLT'95), held March 13–15, 1995, in Barcelona, Spain, consolidated a new series of conferences aimed at fundamental studies of all computational aspects of artificial and natural learning systems. The previous and inaugural European Conference on Computational Learning Theory was held December 20–22, 1993, at Royal Holloway, University of London. (At the first EuroCOLT, preliminary abstracts were distributed at the meeting and the final proceedings was published afterwards as *Proceedings of the First European Conference on Computational Learning Theory: EuroCOLT'93*, J. Shawe-Taylor and M. Anthony, Eds., Oxford University Press, 1994.)

Continuation of the event is supervised by the EuroCOLT Steering Committee, consisting of: M. Anthony (LSE, Univ. London, UK), R. Gavaldà (UPC, Barcelona), W. Maass (TU Graz, Austria), J. Shawe-Taylor (RHBNC, Univ. London, UK), H.-U. Simon (Univ. Dortmund, Germany), P. Vitányi (CWI & Univ. Amsterdam).

The topics discussed in these meetings potentially cover all aspects of analysis of learning algorithms and the theory of machine learning, including artificial and biological neural networks, genetic and evolutionary algorithms, robotics, pattern recognition, inductive logic programming, inductive inference, information theory, decision theory, Bayesian/MDL estimation, statistical physics, and cryptography. Experimental results are welcome, but are expected to be supported by theoretical analysis. In response to our Call for Papers, 46 full draft papers in these areas were submitted by September 21, 1994. Following three rounds of email meetings of the Program Committee, on October 31, 1994, in Amsterdam, 28 submissions were selected for presentation at the conference. In addition, R.J. Solomonoff, J. Rissanen, and A. Macintyre were invited to give lectures and contribute a written version to these proceedings.
The Program Committee for EuroCOLT'95 consisted of:

M. Anthony (LSE, Univ. London, London, United Kingdom);
E. Baum (NEC Research Inst., Princeton, USA);
N. Cesa-Bianchi (Univ. Milano, Milan, Italy);
J. Koza (Stanford Univ., Palo Alto, USA);
M. Li (Univ. Waterloo, Waterloo, Canada);
S. Muggleton (Oxford Univ., Oxford, United Kingdom);
W. Maass (TU Graz, Graz, Austria);
J. Rissanen (IBM Almaden Research Center, Almaden, USA);
H.-U. Simon (Univ. Dortmund, Dortmund, Germany);
K. Yamanishi (NEC Research, Princeton, USA);
L. Valiant (Harvard Univ., Cambridge, USA);
P. Vitányi (CWI & Univ. Amsterdam, Amsterdam, Netherlands, chair);
R. Freivalds (Univ. Riga, Riga, Latvia).

The Local Arrangements Chairs were:

Ricard Gavaldà (UPC, Barcelona, Spain);
Felipe Cucker (Univ. Pompeu Fabra, Barcelona, Spain).

The EuroCOLT'95 conference is sponsored by the EATCS, by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, by IFIP through SSGFCS WG 14.2, and by the Universitat Politècnica de Catalunya.

We want to thank everybody who helped to make this meeting possible: the authors for submitting papers, the Program Committee and referees for their effort in composing the program, the Steering Committee, the sponsors, the local organizers, and Springer-Verlag. The Program Committee wishes to thank the following persons, who acted as subreferees for EuroCOLT'95:

Naoki Abe
Peter Bartlett
Dan Boneh
Harry Buhrman
Zhixiang Chen
Scott Decatur
Claudio Ferretti
Paul Fischer
Lee Giles
Adam Grove
Peter Grünwald
Leonid Gurvits
Tom Hancock
Sanjay Jain
Tao Jiang
Michael Kearns
Joe Kilian
Pascal Koiran
Kevin Lang
Nick Littlestone
Jeroen van Maanen
David Page
Stefan Pölt
Dan Roth
Carl Smith
John Tromp
Peter Yianilos

Amsterdam, January 1995

Paul Vitányi

Table of Contents

SESSION 4
Chair: Ricard Gavaldà

SESSION 5
Chair: Ming Li

SESSION 6
Chair: Jorma Rissanen

SESSION 7
Chair: Hans-Ulrich Simon

SESSION 8
Chair: Wolfgang Maass

SESSION 9
Chair: Kenji Yamanishi

SESSION 10
Chair: Martin Anthony

Editor's Foreword

Advances in algorithmics, computational feasibility, and computer technology, have caused the emergence of learning algorithms in a plethora of computational learning models. This approach can be contrasted with the traditional work in Artificial Intelligence which is logic and semantics based, rather than computation and statistics based.

For industrial applications, it is essential that a unified theory be developed and that techniques be identified to translate this theory into practical applications. For example, it has become apparent that the approach of knowledge-based expert systems, because of the sheer size and complexity of the data, has reached the functional limits of being directly programmed by entering the extracted wisdom of the experts as a full-fledged object (rather like Pallas Athena springing fully armored from the head of Zeus). Future such systems need to learn directly from the experts or experience.

Recent developments in recursion-theory based theories of learning, Bayesian learning, MDL or minimum description length, Kolmogorov (descriptional) complexity, probably approximately correct (PAC) learning, and mathematical approaches to artificial neural learning and genetic learning are emerging as a new general discipline of mathematical foundations underlying practical learning by machines. This has led to the formation of numerous university research groups and commercial laboratory groups, to the 'Annual ACM Conference on Computational Learning Theory (COLT)' (1987–) and other meetings, and to the 'Machine Learning' journal and other more specialized journals. The second EuroCOLT starts a new thread in the Springer-Verlag LNAI series. In these proceedings the invited papers stress historical foundations of the subject and long-term trends while the contributed papers often present probing cutting edge research.

Forecasting and Prediction in On-Line Learning

In Session 1 Ray Solomonoff in his invited paper gives a detailed account of his discovery of algorithmic probability, and as a side product some form of Kolmogorov complexity. These notions are at the heart of all learning: it is always a matter of compression of observations or data into a compact theory. Such compression is only possible if there is a regularity in the data which can be used to compress. Roughly speaking, there is something to be learned from a body of data if and only if that data possesses a regularity which allows us to compress it. This regularity then can be viewed as a theory explaining (or a model for) the data. It is a general feature which is difficult to formally express that the more one compresses the data, the more the resulting theory will generalize and predict. This is the road to true induction: generalizing from the particular to the general. Essentially, Solomonoff's method will forecast all probabilistic phenomena with a constant bounded expected cumulative error.

Solomonoff's result can be viewed as an absolute and objective form of applying Occam's razor 'entities should not be multiplied beyond necessity' by identifying 'simplest' theories with those having the 'shortest effective descriptions', that is, least Kolmogorov complexity.

Since the Kolmogorov complexity of an object is a recursively invariant property, this approach (seemingly casually) also resolves one of the central problems in philosophy, namely to find an objective basis for both the possibility of induction at all and the way to do it.

Unfortunately, among other things, algorithmic probability is not computable. Nonetheless, Solomonoff's approach has proven fundamental in the sense that both Gold-style learning (sometimes called by the general name 'inductive inference') and statistical inference methods like the minimum description length principle (minimum message length principle) are a form of computable approximation to Solomonoff's procedure, as we have argued elsewhere [*JCSS*, 44:2(1992), 343-384]. (A general overview of algorithmic probability and Kolmogorov complexity is M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer-Verlag, New York, 1993.)

The next two papers in Session 1 essentially follow up on Solomonoff's original approach in that they investigate strategies to minimize forecasting errors with respect to an optimal clairvoyant strategy. They estimate the loss with respect to discrepancy between prediction and outcome, which essentially means that one deals with deterministic phenomena. The paper by Freund and Schapire will be especially valuable at the race track, and, moreover, gives (in its particular context) a technique to boost a marginal forecasting strategy into an accurate one. Ben-David, Kushilevitz, and Mansour show to what extent the order in which the data are presented influences the amount of mistake the forecaster makes.

Seminal work by V.G. Vovk [*Inform. Comput.*, 96(1992), 245-277] formulated a more tractable strategy of forecasting of phenomena in the spirit of Solomonoff. In Session 2 this style of work is represented by Haussler *et al.* who continue a series of papers following Vovk's lead. They determine (in particular settings) the correct value of the worst-case total loss incurred by the forecasting algorithm for a variety of loss functions such as square loss, logarithmic loss, and Hellinger loss. They show by establishing a lower bound that this value is asymptotically tight.

Yamanishi introduces a new family of deterministic and stochastic on-line predictors. He approaches the problem of on-line learning by iterative prediction-outcome-adapt predictor cycles in a more general stochastic setting. The predicted phenomenon is viewed as a parametric probabilistic model, and the algorithm iteratively estimates the probabilistic model using maximum likelihood and constructs an optimal predictor minimizing the average loss with respect to the probabilistic model. The next outcome is predicted by this optimal predictor. This method allows us to analyze upper bounds on expected instantaneous and cumulative losses with respect to a large family of loss functions and conditions on the parametric probabilistic models. Its novelty is that the method designs

and analyzes on-line prediction algorithms with respect to *expected loss bounds* for large classes of loss functions simultaneously.

Inductive Inference and Recursion Theoretic Learning

There is a strong tradition based on recursion theory in the theoretical learning community which goes back at least to Gold's paradigm of 'learning in the limit' and 'learning by enumeration'. We have argued elsewhere [*JCSS*, 44:2(1992), 343-384] that if not historically, then at least logically, Gold's paradigm can be viewed as a special case of Solomonoff's approach. In learning by enumeration we start with an enumeration of hypotheses and eliminate inconsistent hypotheses from the beginning of the list up to the first hypothesis consistent with the evidence received up till now. This idea can be elaborated with all sorts of embellishments. For example, suppose the learner outputs a sequence of hypotheses in the learning process. Each time the learner outputs a hypothesis different from the previously output hypothesis he experiences a 'mind change'. By procrastination the learner can possibly reduce the number of mind changes. To formalize inference with procrastination constructive ordinals are used. Previous research left open the question how the notation of these ordinals can influence the power of inference. This question is investigated in Section 3 by Ambainis. In the same section Lange and Zeugmann consider the relation between monotonicity demands on the sequence of output hypotheses (formalized notions that the consecutive output hypotheses are better and better generalizations) versus number of mind changes.

In Section 4 Case *et al.* introduce the idea of learning recursive programs in the limit from not only input/output examples of the target computable function but also varieties of approximate information about the target function (like frequency counts). Freivalds *et al.* introduces a more structured abstract view of the interrelations between learnability classes by introducing natural notions of reducing one class to another. This is done in the setting of inductive inference much like the traditional field of so-called 'structures in (computational) complexity theory' where such notions are studied with respect to time- and space-bounded complexity classes of problems like the ubiquitous classes P and NP. They establish hierarchies of more and more difficult-to-learn concepts and completeness results. The results indicate (the authors state) that the complexity notion captured in their novel approach differs dramatically from the traditional study of computational complexities of the algorithms performing learning tasks. This introductory paper is followed immediately by a further exploration of the new concepts by Jain and Sharma.

Another aspect of inductive inference is to identify a 'minimal length' or almost minimal length program for a target function. This depends on the expression of the programs, which can be done via Gödel numbering. Freivalds and Jain investigate properties of minimal identification in Kolmogorov numberings—in some sense optimal numberings. Relating back to another fundamental notion due to Kolmogorov, in Section 3 Balcázar *et al.* show that circuit expressions of

low time-bounded Kolmogorov complexity are polynomial time learnable from membership queries in the presence of an NP oracle.

In Section 6, the paper by Martin and Osherson establishes the interesting and seemingly paradoxical fact that the successful strategy for a learner who is constrained to effective computation (like a Turing machine) may not be to compute (and act on) the conditional probabilities of hypotheses given the data received so far (even if he could do so). In contrast, for learners using some non-effective computational capabilities the successful strategy is always to compute the conditional probabilities. This counterintuitive result is couched in very abstract terms. It would be interesting to see whether it can be translated to more directly appealing exemplary 'paradoxes'.

In Section 10 Koshiba shows how to identify typed pattern languages in the limit from positive examples and related results.

Stochastic Phenomena and MDL Learning

Algorithmic probability in the sense of Solomonoff, and shortest effective description length in the sense of Kolmogorov, have influenced approaches in statistics based inductive inference. Rather than trying to estimate a 'true' data generating distribution, whose existence may be entirely mythical, as in traditional statistics, one considers the (almost) objective notion of minimizing the code length of the data given a model (hypothesis) class and the code length of the model. This is the so-called minimum description length (MDL) principle. Replacing the troublesome notion of applicable 'probability' (variously identified with frequencies of mass events, or initial belief of the learner) with a more objective and tangible notion as 'code length' one hopes to escape epistemological quicksand. To avoid noncomputability we settle for codes whose lengths are some computable upper bound on the noncomputable Kolmogorov complexity. In the form of 'minimum message length' this idea was invented separately by C.S. Wallace and J. Rissanen, the last of whom coined the name MDL. Rissanen also formulated the notion of 'stochastic complexity', an approximation of Kolmogorov complexity relative to the constraints of a given model class as follows. The stochastic complexity is a lower bound on the mean of the code lengths and almost surely on the code length induced by each individual model, the possible exceptions being restricted to cases when the string is generated by models in a subset of the model class of measure zero. In an expository invited paper starting Session 5, Jorma Rissanen explains the notions involved, including the most recent explicit expression of stochastic complexity in a satisfactory form. He provides applications to problems arising in machine learning such as giving improved designs for MDL decision trees.

In Section 6, remarkably, Vovk shows one does not always need to be bothered by the noncomputability of the Kolmogorov complexity. (Rather, as is more appropriate here, the prefix-free version of it associated with the names of Levin and Chaitin.) He shows that for some standard statistical models one can find computable minimum description length estimators in the sense of Kolmogorov

complexity. The paper by Kilpeläinen *et al.* gives an example of an MDL learning algorithm (for pattern languages).

PAC Learning and Query Learning

The Solomonoff procedure (apart from being noneffective as well) and identification by enumeration have an unpractical drawback. The learning program never knows when to stop. Moreover, in many cases it is required that all examples must appear eventually, although this is not required for some concept classes (which can be uniquely determined by a finite set of examples). Given a new example, finding the first or minimum rule consistent with the data usually requires time which is exponential in the size of the rule.

The model of distribution-free learning introduced in 1984 by L.G. Valiant [*CACM*, 27(1984), 1134-1142] had the purpose of describing 'feasible' learning. According to common views in the theory of computation, 'feasibility' is identified with the requirement that the learning algorithm should run in polynomial time and use a polynomial number of examples. This requirement implies that not all examples can turn up. Hence, it is impossible to infer a concept precisely. This means that we can only hope to learn the concept approximately. But we also have to account for where the examples originate. We could be presented with unrepresentative examples for the concept to be inferred. It seems reasonable to assume that the examples are drawn randomly from a sample space according to a probability distribution. The approximation between the target concept and the learned concept can then be expressed in terms of the total probability of the set of examples on which the two concepts disagree: the expected classification error.

As long as the learner does not see all examples, he can always err badly. The key is to make sure that happens with very small probability. In practice, the speed of learning is realized only because learning is just 'approximately correct' with 'high probability'. For example, after a child is given several sample apples, he likely forms an approximately correct concept of apple. In this model the learner passively gets examples drawn from a probability distribution. This learning model has been termed 'probably approximately correct (PAC) learning' by D. Angluin.

In Section 2 Ben-David and Lindenbaum essentially analyze a notion of cluster learning and relate it to PAC learning. Their task is to learn unsupervised from unlabeled examples. The only information about their membership is indirectly given to the student through the sampling distribution. This means that one infers information about the underlying distribution from the random examples it generates. They develop a learning model for such problems, and show that a class is learnable in this setting if and only if the Vapnik-Chervonenkis dimension is finite.

In Section 5 Anthony and Bartlett consider PAC learning of real-valued functions from random argument-value pairs (interpolations). This also yields applications to learning functions in the presence of malicious noise. Also in Section 5, Gurvits and Koiran show that the scale-sensitive version of the pseudo-dimension

has proved crucial in an understanding of function and p-concept learnability, and the authors provide a construction of function classes having bounded scale-sensitive dimension. Of particular relevance is the material where function classes corresponding to simple neural nets are discussed. These are among the first non-trivial results estimating scale-sensitive dimensions of interesting concept classes like neural networks and function classes considered by A. Barron.

In Section 7 Schuurmans resolves the problem that according to a known mathematical result the learning curve of the smallest expected classification error in the worst case converges like $1/t$ for a training sample size t, while in experimental settings often an exponential convergence like $\exp(-t)$ can be observed. This discrepancy is shown to occur due to the fact that for finite concept classes consistent learners can expect exponential convergence, while for continuous concepts no learner can exhibit subrational (below $1/t$) convergence in all circumstances. In Section 9 Golea analyzes the learning curve for the average case of a learning algorithm for μ-DNF formulas under the fixed uniform distribution (rather than distribution-free). He obtains a square average sampling complexity, a cube root improvement over the known PAC analysis result for the same problem. Empirical evidence is also provided.

A related model is D. Angluin's notion of learning by membership and/or equivalence queries. Here there is no distribution the examples come from, but the learner can ask a teacher questions. Such a question can either be 'is this example a positive or a negative example', resulting in a correct identification of the type by the teacher, or the question can be 'is this hypothesis the looked-for target concept', resulting either in the answer 'yes' if it is or in a counterexample if the answer is 'no'. The first type is called a 'membership query' and the second type is called an 'equivalence query'.

In Section 8 Ameur demonstrates an equivalence query learning algorithm which learns axis parallel boxes in less space than other known algorithms at the cost of a moderate increase in time complexity. Also for equivalence queries, Simon shows that decision lists and decision trees of bounded rank are polynomially learnable.

In Section 9 Shamir and Shwartzman introduce yet another learning model. This time we do not poll for example membership queries, nor do we query whether our current choice of hypothesis is the target concept (obtaining a counterexample if it is not), the equivalence query model, but now we consider statistical queries. That is, we can ask for the expected value of expressions involving the unknown target concept. The paper extends Kearns *et al.*'s statistical query model in order to be able to cast known efficient PAC learning algorithms in statistical query format.

In Section 10 Bergadano and Varrecchio show how to learn certain extensions of finite automata exactly in polynomial time using equivalence queries. Also in Section 10, Domingo and Lavín show relations (with respect to alternations) between membership queries and equivalence queries to learn certain classes of restricted context-free languages.

Artificial Neural Network Learning

Artificial neural networks are widely used in practice and are starting to be thoroughly investigated as to the fundamental mathematics underpinning their workings. Relatively little was known about the computational power of polynomially sized feed-forward neural networks with smooth (sigmoid) gate functions or the tradeoff between size and number of layers in such networks. In his invited paper in Section 9, Angus Macintyre (with co-author Marek Karpinski) surveys the area and describes their recent result of a polynomial bound on the Vapnik Chervonenkis dimension of sigmoid neural networks (thresholded at the output).

In Section 7 Muselli revisits a fundamental problem in training threshold node circuits (perceptrons). When the concept involved is not linearly separable, then there is no assignment of weights which allow the perceptron to correctly identify all examples. Moreover, while for linearly separable concepts there are known algorithms which converge fast, for nonlinearly separable concepts there does not seem to be any other algorithm that converges to an optimal separation (minimal number of errors) than the so-called 'pocket algorithm'. Hence this algorithm may play a basic role also in training multinode neural networks. The current paper analyzes more carefully than was done before the conditions under which the pocket algorithm converges (restricting the range of validity of the original statement of this result) and eliminates some formal problems with the original proof.

Also in Section 7 van den Berg and Bioch consider general stochastic binary Hopfield models from the viewpoint of statistical mechanics, and derive some theorems expressing the amount of free energy of such networks.

Learning from Patterns

In Section 10 Brāzma exhibits simple procedures to learn regular expressions from example patterns which are 'good' in the particular sense described by the author.

Paul Vitányi