Editors' Introduction

This volume contains all the papers presented at the *Seventh International Workshop on Algorithmic Learning Theory* held at the Coogee Holiday Inn, Sydney during October 23–25, 1996. The program consisted of 16 long papers and 8 short papers in addition to four invited papers by distinguished researchers representing different aspects of algorithmic learning.

In this introductory note, we give the reader a tour of each of the sessions and the papers presented in them. We attempt to give an informal overview rather than a technical critic. We do not, however, attempt to classify the research reported into one of the many branches of algorithmic learning, but do try to set the context in which these papers appear, and occasionally point the reader to related work.

Session 1

The opening session of the workshop contains an invited presentation by Leslie Valiant. In his 1994 book, *Circuits of the Mind*[1], he introduced a framework for investigating effects of biological constraints on issues of evaluation and learning. In the original exposition of the framework, some central questions about the management of high descriptional complexity were left unresolved. In his invited paper in this volume, he considers three sources of such difficulty: conflict resolution, implementation of complex strategy modules, and robustness. He argues that the conceptual tool of "prioritized threshold circuits" can be used to manage significant aspects of all three sources of complexity. He concludes with a discussion of the implications for Artificial Intelligence.

Session 2

Session 2 consists of three long papers and two short papers. The topics covered in this session include learnability of Boolean formulas, on-line learning of functions over reals, and empirical results on algorithms for inference of minimum-size finite-state machines.

Learnability of DNF formulas has been a longstanding problem in algorithmic learning theory. Bshouty[2] initiated the approach of decomposing Boolean functions into monotone functions to tackle this problem. In this volume, Takimoto, Sakai, and Maruoka propose a new kind of decomposition, exclusive-or expansion based on monotone DNF formulas, to investigate the learnability of this class in Angluin's query learning model. They show that any Boolean function can be represented as a formula in this class. Another interesting property of this representation is that every Boolean function has a unique reduced form. They derive various algorithms for learning such formulas.

Long addresses the complexity of learning smooth functions on the reals. These functions capture the idea that similar inputs yield similar outputs. The property of smoothness is achieved by bounding norms of a function's derivative. A nice tight best-possible bound on the worst-case sum of absolute prediction

[1] L. Valiant, *Circuits of the Mind*, Oxford University Press, 1994.

[2] N.H. Bshouty, "Exact learning via the monotone theory", FOCS 1993.

errors is derived in the mistake-bound learning model adopted for continuous-valued functions.

Ordered binary decision diagrams (OBDDs) are a useful representation of Boolean functions, and their learnability has recently been studied in both the query learning model and the PAC model. In this volume, Nakamura studies the exact learnability of *bounded-width* OBDDs. The present study differs from the earlier study by Ergün et al.[3] in that Ergün et al. assume the ordering of variables in the target OBDD to be given, whereas Nakamura studies the more difficult case where the ordering is unknown. An algorithm using only $O(n^3)$ equivalence queries and another algorithm using $O(n)$ proper equivalence queries and $O(n^2)$ membership queries are given for learning width-2 OBDDs. The author also shows that no polynomial time algorithm can learn width-3 OBDDs from proper equivalence queries alone.

In their short paper, Büning and Lettman observe that since the exact learning model requires identification of the target formula up to logical equivalence, the learning algorithm may be used to optimize the representation of a formula with respect to deducibility. They modify the standard query model to include clause queries, which determine whether a given clause is a consequence of the target formula, to show that simple representations can be efficiently learned for some classes that are not learnable with membership and equivalence queries.

The short paper by Oliveira and Edwards investigates the problem of inferring finite-state machines with the minimum number of states consistent with a given training set. The paper has an empirical flavor and compares four different algorithms for this task.

Session 3

This session consists of a invited presentation by Paul Vitányi and two short papers that address aspects of decision-tree construction.

In his invited paper "Genetic fitness optimization using rapidly mixing Markov chains", Paul Vitányi argues that the use of unbounded or exponential population sizes in the performance analysis of genetic algorithms is not directly applicable to practical problems. He proposes a notion of highly probable fitness optimization through evolutionary runs on small-size populations in a very general setting. He then uses recent results about rapidly mixing Markov chains to show that under suitable conditions, the proposed method efficiently finds optimal programs with a probability of almost 1. The analysis presented in the paper justifies the empirical practice for certain genetic computations of using a large number of short runs in preference to one long run. The theory developed is illustrated in the context of a toy example.

The first of the two short papers by Baxter and Oliver addresses the topic of optimal segmentation, an important issue in decision-tree construction. They derive minimum message length expressions for region boundaries and observe that the message length cost is dependent on the noise of the data in the separated regions and on the degree of separation of the two regions.

[3] F. Ergün, R.S. Kumar, and R. Rubinfeld, "On Learning Bounded-Width Branching programs", COLT'95.

The second short paper in this session, by Pearson and Smith, describes a number of heuristics aimed at reducing the "complexity" of built trees with respect to disjunctive concepts. They provide empirical evidence that certain combinations of their heuristics perform better or as well as conventional partitioning techniques.

Session 4

The fourth session consists of two long and two short papers addressing issues in exact-learnability of algebraic concepts, learnability of counter automata, and feature selection in machine learning.

Arvind and Vinodchandran address the issue of comparing two concept classes that are both exactly learnable with polynomially many equivalence queries. In particular, they consider the scenario in which a class $\mathcal{A}$ is easier to exactly learn than class $\mathcal{B}$ because the full power of equivalence queries may not be needed to learn $\mathcal{A}$ but is required to exactly learn $\mathcal{B}$. They propose the notion of *teaching assistants* to develop a finer criterion for classifying the complexity of exact learning. They then apply this machinery to analyze the complexity of exact-learning classes of permutation groups and linear spaces over finite fields.

Fahmy and Roos give an efficient algorithm that learns deterministic real-time two-counter automata (R2CA) by automata decomposition. Their learning algorithm adapts Angluin's algorithm[4] for learning regular sets. It uses a minimally adequate teacher that provides the shortest counterexample and is capable of answering equivalence queries for R2CA. This work is an extension of the authors' earlier work reported in ALT'95 in which they used the decomposition method of Hartmanis and Stearns[5] to derive an efficient algorithm for learning deterministic real-time one-counter automata.

The two short papers in this session address the issue of feature selection in machine learning. The first of these papers, by Lavrač, Gamberger, and Turney, investigates whether it is possible to detect any information contained in the training examples and the background knowledge that facilitates elimination of irrelevant features in the preprocessing phase before learning takes place. They present a case study of a hybrid genetic algorithm that supports elimination of irrelevant features to improve the efficiency of learning.

The second short paper, by Piramuthu, proposes a new feature selection method based on the "Blurring" measure[6]. An empirical comparison of this method against information-theoretic measures and stepwise multiple regression is reported.

Session 5

This session begins with an invited presentation by J.R. Quinlan and includes one long paper and two short papers.

---

[4] D. Angluin, Learning regular sets from queries and counter examples, *Information and Computation*, Vol. 75, 1987.

[5] J. Hartmanis and R.E. Stearns, *Algebraic Structure Theory of Sequential Machines*, Prentice Hall, 1966.

[6] L. Rendell and H. Raghavan, "Improving the design of induction methods by analyzing algorithm functionality and data-based concept complexity", IJCAI, 1993.

The theoretical technique of *boosting*, originally introduced by Schapire[7], has recently found application in practical machine learning. For instance, **AdaBoost**, a version of the boosting algorithm developed by Freund and Schapire[8], has been shown to improve the accuracy of classifier-learning systems like C4.5. The invited paper by J.R. Quinlan in this volume reports on empirical results of boosting on a first-order learning system FFOIL. The results reported suggest that although the improvement is not as marked as that for propositional-level learning systems, boosting is likely to be helpful in the context of first-order learning systems, too.

Bārzdiņš and Sarkans consider the problem of inducing of functions not only from input/output examples but also from knowledge about the syntactic description of the function and about evaluation of the function. They take functional description to be an expression over some fixed signature. They then argue that various kinds of hypothetical knowledge about the unknown function can be described in terms of an attribute grammar with synthesized attributes. They then show that under certain conditions it is possible to efficiently enumerate the corresponding language without considering strings that do not belong to it. This efficient enumeration of the hypothesis space is likely to turn out to be useful in the design of practical learning systems.

The short paper by Martin and Vrain addresses the problem of why most inductive logic programming systems restrict their hypothesis space to not include function symbols. They propose a method based on the utility of a function symbol to discriminate between positive and negative examples to include function symbols in inductive logic programming. They model their ideas in the framework of constraint logic programming and show that their algorithm is capable of learning some constraint logic programs.

Sugimoto, Hirata, and Ishizaka consider the problem of learning translations based on a dictionary. They introduce a restricted elementary formal system to formalize learning translations as extraction of binary relations over strings from given positive and negative examples based on information available from a dictionary. They describe a bottom-up algorithm that is capable of learning translations over context-free languages.

Session 6

The three long papers in this session address generalization issues in inductive logic programming, noise handling in machine learning, and application of the Minimum Message Length principle to a statistical problem.

Lu and Arima argue that the set inclusion ordering on the success set of logic programs is a more useful notion of generalization in the context of inductive logic programming than the traditional approaches based on clausal θ-subsumption and clausal logical implication. They propose three kinds of generalization orderings between programs. Based on these orderings, they present

[7] R.E. Schapire, The strength of weak learnability, *Machine Learning*, Vol. 5, 1990

[8] Y. Freund and R.E. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting", EuroCOLT'95, Springer-Verlag, 1995 (see also Freund and Schapire, "Experiments with a new boosting algorithm", ICML'96).

a set of generalization rules borrowed from the unfold/fold program transformation method. They also discuss strategies for application of these rules.

Traditional techniques of handling noise in machine learning take the approach of the learner trying to avoid overfitting the noisy example set. Gamberger, Lavrač, and Džeroski propose separation of noise detection and hypothesis formation so that noisy examples do not influence the construction of the hypothesis. They describe a simple compression-based measure to detect noise, and propose a technique that eliminates the noisy examples and then constructs the hypothesis from the remaining examples. They demonstrate the potential of their approach by applying it to a problem of early diagnosis of rheumatic diseases.

The spherical Fischer distribution is an important distribution in statistical data analysis. Dowe, Oliver, and Wallace apply the information-theoretic Minimum Message Length (MML) principle to the problem of estimating the concentration parameter of spherical Fischer distributions. They carry out a complete calculation of the parameter for a special case, and report that simulation results show the MML estimator to compare favorably with the classical approaches of Maximum Likelihood and marginal Maximum Likelihood.

Session 7 was devoted to the *Pacific Rim Knowledge Acquisition Workshop* invited talk by William Clancey.

Session 8 consisted of a poster session for the short papers.

Session 9

Motivated by the observation that it is easier to discard incorrect hypotheses rather than discover a correct one, Freivalds, Karpinski, and Smith[9] have recently studied identification in the limit with an interesting twist. In this model, a learning machine, instead of converging to a correct hypothesis, eliminates hypotheses. The hypothesis with the minimal index that never gets erased is assumed to be the hypothesis that the machine stabilizes to. The two papers in this session deal with this notion, referred to as co-learnability or learning by erasing.

In the first paper, Lange, Wiehagen, and Zeugmann extend the notion of co-learnability in several directions in the context of learning indexed families of recursive languages. They consider both learning from texts (positive data) and from informants (both positive and negative data). They give characterizations of ensuing learning classes in addition to comparing them to the traditional learning classes in the literature. An interesting observation of their study is that the process of elimination cannot be restricted to only incorrect hypotheses for achieving its full learning ability. This result reinforces that a learner has to occasionally be allowed to discount correct hypotheses if it is to achieve its full learning potential.

The second paper in this session on co-learning, by Jain, Kinber, and Wiehagen, studies co-learnability of minimal programs. They show that co-learning of minimal programs is significantly weaker than learning of minimal programs even

[9] R. Freivalds, M. Karpinski, and C.H. Smith, "Co-learning of total recursive functions", COLT'94.

in Gödel numberings. To address this weakness, they consider extensions of the notion of co-learnability. They study various relationships between the resulting notions and also explore co-learning in the natural Kolmogorov numberings. A number of additional issues are addressed and some of the proof techniques are elegant.

Session 10

This session included an invited talk by Takeshi Shinohara and two long papers. One of the long papers builds a bridge between logic programming and inductive inference, the other considers vacillatory and behaviorally correct identification in the presence of noise.

Shinohara and Arimura in their invited contribution consider inductive inference of unbounded unions of pattern languages from positive data. Angluin[10] had initially shown that pattern languages can be identified in the limit from positive data. Shinohara[11] later observed that the class of pattern languages is not closed under union, and showed that the union of up to two pattern languages is identifiable in the limit from positive data. Wright[12] later generalized this result to show that for any fixed k, the union of k pattern languages is identifiable in the limit from positive data. However, in many applications like genome informatics such an upper bound is not known in advance. Shinohara and Arimura point out that the union of an unbounded, but finite, number of pattern languages is not identifiable in the limit from positive data. They also show that this unidentifiability result does not change even if attention is restricted to proper patterns. They then work out some additional restrictions under which unbounded unions are identifiable. On the way they notice some interesting links between the identification of extended pattern languages and the present problem of unbounded unions.

Krishna Rao describes a class of logic programs identifiable in the limit from positive data. This class is a generalization of an earlier class shown to be inferable by Arimura and Shinohara[13]. The author introduces a class of linearly moded logic programs which is a proper superclass of linearly covering programs considered by Arimura and Shinohara. The inferability is established by showing that the class is an indexed family and has bounded finite thickness.

Stephan introduced a model of noise in his ALT'95 paper in which correct information appears infinitely often and incorrect information appears only finitely often. In this volume, Case, Jain, and Stephan study vacillatory and behaviorally correct identification of languages in Stephan's noise model both for texts (positive data only) and for informants (positive and negative data). It is well known

[10] D. Angluin, Inductive inference of formal languages from positive data, *Inf. and Control*, Vol. 45, 1980.

[11] T. Shinohara, Inferring unions of two pattern languages, *Bull. of Inform. and Cybernetics*, Vol. 20, 1983.

[12] K. Wright, "Identification of unions of languages drawn from an identifiable class", COLT'89.

[13] H. Arimura and T. Shinohara, "Inductive inference of Prolog programs with linear data dependency from positive data", Proc. Information Modelling and Knowledge Bases, IOS Press.

that vacillatory identification from informants coincides with identification in the limit from informants when no inaccuracies are present in the data. Case, Jain, and Stephan show that in noisy environments, this is not the case. They exhibit various noisy-data hierarchies. They also show that noisy behaviorally correct learning obeys a very strong "subset principle".

Session 11

The last session of the workshop consists of three presentations on inductive inference.

Motivated by the success of algebraic approaches to studying natural transformations under which an object remains invariant, especially in mathematics and theoretical computer science, Ambainis and Freivalds begin a study of transformations that preserve learnability in the inductive inference paradigm. They model natural transformations by general recursive operators. They show that there are transformations that preserve finite identifiability but which do not preserve identification in the limit, and vice versa. They also show that transformations preserving identification with up to i mind changes always preserve identification with up to j mind changes where $j < i$.

Most studies of identification in the limit with errors in the converged hypothesis have considered only finitely many errors. This approach has often been criticized on the ground that it does not make sense to study a finite number of errors when the concepts being learned are infinite. Royer[14] and Smith and Velauthapillai[15] have considered approaches where the final converged hypothesis makes an infinite number of errors, but with "low density". Viksna, in his paper in this volume, proposes an alternative approach to measuring the error set by defining a "small" error set to be one with zero measure. He studies relations between classes of functions identifiable up to "small" error sets for different choices of measure. Then he goes on to study this notion of error in the framework of probabilistic identification in the limit.

Grieser investigates inductive inference machines that are capable of reflecting about their own limitations. The notion of reflecting learners was first considered by Jantke in his ALT'95 paper, where he showed that for identification in the limit, reflection is possible if and only if it is not needed. Griesser shows that such is not the case for finite identification and considerably extends Jantke's results. This work is a step in the direction of developing systems that are capable of assessing their competence. The paper also introduces the notion of therapy to handle situations in which a learner signals its inability to learn.

At the end of this session, ALT'96 concluded.

Fukuoka, August 1996 — Setsuo Arikawa
Sydney, August 1996 — Arun Sharma

---

[14] J. Royer, Inductive inference of approximations, *Information and Computation*, Vol. 70, 1986.

[15] C.H. Smith and M. Velauthapillai, On the inference of approximate programs, *Theoretical Computer Science*, Vol. 77, 1990.