A Book Review: *A Theory of Learning and Generalization* by M. Vidyasagar

Jiří Šíma *

In general, heuristics used in the field of artificial intelligence have often been inspired by human intuition. Although most of these methods produce useful results in practical applications, their rigorous mathematical analysis which may answer important questions such as how a particular heuristic solution can be verified?, why a given algorithm sometimes fails?, under which conditions a method can safely be applied?, etc., is usually omitted. This is probably caused by the fact that the respective theoretical analysis is usually very difficult and there are no simple answers to the above-mentioned questions. In addition, practitioners rarely communicate their issues to theoretical computer scientists who deal with their own 'interesting' problems.

A similar situation was in the field of machine learning, especially in the artificial neural networks that are frequently exploited as black-box tools. Here, several intuitive questions have appeared such as which concepts or functions are, in principle or efficiently, learnable within particular models?, how many training data is required for successful generalization?, etc. Therefore, in his seminal 1984 paper [10], L. G. Valiant attempted to formalize the phenomenon of learning and generalization by introducing the well-known *PAC* (**P**robably **A**pproximately **C**orrect) model which gave rise to a new informatics branch—the *computational learning theory*. Namely, for *samples* of an unknown target concept (from a given concept class) which are presented in accordance with an unknown fixed distribution, a PAC polynomial-time learning algorithm produces a hypothesis that, with a desired *confidence*, approximates the target concept within a prescribed *accuracy* in terms of the underlying probability measure. Along these lines many important and relevant results have been achieved, and by now, it is widely accepted that the PAC learning formulation presents a mathematically rigorous basis for studying the intuitive idea that 'neural networks can generalize'. Actually, neural network models serve here as the application or testing area of the general learning theory as well as a source of motivations for further research.

Recently, several excellent monographs or survey papers on learning theory, such as [1, 2, 5, 6, 7, 9], appeared which stress the *computational* aspects of learning and generalization. This is now completed by the present book *A Theory of Learning and Generalization* (*With Applications to Neural Networks and Control Systems*) by Mathukumalli Vidyasagar (Director of the Center for Artificial Intelligence and Robotics, Bangalore, India) which was published within the *Communications and Control Engineering Series* in *Springer-Verlag*, London, 1997, and this work emphasizes the *statistical* ideas behind the PAC model. Thus the main objective of this book is to bring out clearly the connection between the (PAC) learning theory and some of the fundamental results in the *theory of empirical processes*.

A parallel development in the theory of empirical processes which was initiated by J. Bernoulli more than two centuries ago, was to have a profound impact on

*Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, e-mail: sima@uivt.cas.cz. Research supported by GA ČR Grant No. 201/98/0717.

the learning theory. The most influential work was, in particular, the 1971 paper [11] by V. N. Vapnik and A. Y. Chervonenkis who gave necessary and sufficient conditions for empirical estimates of the probability measures of a family of sets to converge to their true values, as the number of samples approaches infinity. In these conditions, they introduced a combinatorial parameter which has since come to be known as the *Vapnik-Chervonenkis* (*VC-*) *dimension*. Within the learning theory the VC-dimension was exploited by A. Blumer *et al.* [3] whose 1989 paper was apparently the first to make a connection between the PAC model and the theory of empirical processes. Thus, the VC-dimension is exploited to estimate the *sample complexity*, i.e. the number of samples that are necessary for correct generalization. According to the author's opinion '*... this connection remains unexplored, or perhaps merely unexplained to a wide audience. In particular, by reformulating the learning problem as a convergence problem for stochastic processes, it is possible to make a distinction between* information-theoretic *limitations to learning, and* complexity-theoretic *limitations to learning. Roughly speaking, information-based learning theory attempts to study what is learnable* in principle*, whereas complexity-based learning theory attempts to study what is learnable* in practice*.*'

Hence the present book is focused on the exposition of the theory of empirical processes for the use in the PAC learning theory. Besides the introductory Chapters 1 and 2 providing motivations and reviewing some of the basic concepts and results from the probability theory, Chapter 3 recalls the statistical PAC learning problem (excluding the computational aspects such as the complexity of learning algorithm, concept representation, etc.) for concepts and functions and its variants (*probably uniformly approximately correct* and *agnostic* learning) showing their relations to the uniform convergence of empirical means. Chapter 4 defines the VC-dimension which is applicable to problems of the uniform convergence of empirical probabilities and of concept learning, whereas its generalization—the *P-dimension* (*Pollard* or *pseudo-dimension*) is suitable for the problems of the uniform convergence of empirical means and of function learning. In addition, several technical tools to derive simple upper bounds for the VC-dimension of complicated collections of sets, are shown here.

Chapter 5 represents the core of the book where several necessary and sufficient conditions for the uniform convergence of empirical means to their true values as the number of samples approaches infinity, are stated and proved. Two versions of this problem are considered: either the probability measure generating the samples is known and fixed, or it belongs to a known family of probabilities (here the underlying convergence is required to be uniform also with respect to the probability). Then in Chapter 6, these results are applied to problems of concept and function learning when the samples are drawn in accordance with a *known fixed distribution*, in order to achieve various necessary and/or sufficient conditions for a concept or function class to be learnable. For example, these characterizations of statistical learnability include the property of uniform convergence of empirical means or the 'shrinking width' property or the 'finite metric entropy'. Further, Chapter 7 is focused on the other extreme, namely the *distribution-free learning* which means there is a complete absence of any prior knowledge about the probability measure generating the samples. In this case, simple necessary and/or sufficient conditions for learnability involving the VC-dimension or the P-dimension may be derived. The intermediate situation when the learning samples are generated by a probability measure belonging to a family which is neither a singleton set nor the set of all probability measures is studied in Chapter 8. Some learnability conditions are derived for special cases of the underlying probability measure family (e.g. when some metric structure is imposed).

Furthermore, in Chapter 9 also the notion of *efficient* learnability which requires a polynomial-time learning algorithm and the polynomial sample complexity as the accuracy and confidence parameters approach their limits, is inspected. In addition, some variants of the standard learning model are taken here into account. Namely in the *active* learning the learner controls the generation of sample data, e.g. selects an input sample querying whether it belongs to the unknown target concept, or the learning algorithm is permitted to make certain binary queries. Also the *passive* learning with perfect or noisy oracles are considered. By results from Chapter 6 and 7 it appears that the notion of learnability is robust within these variants. Particularly, a concept class is learnable in any one of the three models of learning if and only if it is learnable in all of them—the only difference is the sample complexity, i.e. the number of samples required to achieve a specified level of accuracy might become significantly larger. On the other hand, in case some prior information about the target concept is known, it turns out that the conditions for learnability are qualitatively different.

Finally the presented general results from statistical learning theory are applied to neural networks in Chapter 10 and to control systems in Chapter 11 and several open problems are listed in Chapter 12. In particular, for neural networks the sample complexity (the number of training patterns) in terms of the VC-dimension is estimated in order to achieve the correct generalization. For this purpose polynomial upper and lower bounds on the VC-dimension are derived for multi-layered perceptron networks with threshold and piecewise-polynomial activation functions. This is then generalized for networks with 'Pfaffian' activation functions including the standard sigmoid which is employed in the well-known back-propagation learning algorithm. In addition, a technique of proving the VC-dimension to be finite without having an explicit upper bound is presented. It turns out that there are no *information-theoretic* barriers to learning and generalization in the above-mentioned neural network models with the finite VC-dimension. On the other hand, an example of a neural network with a simple, although artificial, continuous activation function which has an infinite VC-dimension is also shown.

However, there are mainly *complexity-theoretic* barriers to efficient learning in neural networks. This is caused by the fact that the so-called *loading problem* is NP-complete, i.e. it is hard to decide whether there exist weight parameters for a fixed network architecture such that the neural network is consistent with a given training data. Finally, the method of *structural risk minimization* to find an optimal neural network architecture that avoids 'overfitting' is discussed. On the other hand, it seems that the author is an excellent expert in the respective parts of the probability theory rather than in the theory of neural networks whose explanation appears to be incomplete in comparison with the preceding general probabilistic analysis. For example, Problem 12.12 (p. 370) listed among the open problems had been solved [8, 4] before the present book was published.

In my opinion, the present book is a readable exposition of the theory of empirical processes clarifying its connection with the learning theory. A reader can learn from it a deep insight into the statistical aspects of the learning theory. Further, most of the chapters are relatively self-contained and can be studied independently when one is interested only in a particular issue. Moreover, the results are always clearly motivated by illustrative examples and some proofs from literature are simplified. Thus I can recommend the present book to all researchers, university teachers, and graduate students who are interested in the theoretical foundations of learning, neural networks and control systems.

References

[1] Angluin, D.: Computational learning theory: Survey and selected bibliography. In *Proceedings of the 24^{th} Annual ACM Symposium on the Theory of Computing*, pp. 351–369, ACM Press, New York, 1992.

[2] Anthony, M., Biggs, N.: *Computational Learning Theory.* Cambridge University Press, Cambridge, 1992.

[3] Blumer, A, Ehrenfeucht, A., Haussler, D., Warmuth, M. K.: Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**:929–965, 1989.

[4] Hammer, B.: Training a sigmoidal network is difficult. In *Proceedings of ESANN'98*, pp. 255–260, 1998.

[5] Hegedüs, T.: Computational learning theory and neural networks: An introduction. In *Proceedings of the SOFSEM winter school, Milovy (Czech Republic)*, M. Bartošek (editor), pp. 77–95, 1994.

[6] Kearns, M., Vazirani, U.: *Introduction to Computational Learning Theory.* MIT Press, Cambridge, 1994.

[7] Maass, W.: Perspective of current research about the complexity of learning on neural nets. In *Theoretical Advances in Neural Computation and Learning*, V. P. Roychowdhury, K.-Y. Siu, and A. Orlitsky (editors), pp. 295–336, Kluwer, Boston, 1994.

[8] Šíma, J.: Back-propagation is not efficient. *Neural Networks*, **9**:1017–1023, 1996.

[9] Turán, G.: Computational learning theory and neural networks: A survey of selected topics. In *Theoretical Advances in Neural Computation and Learning*, V. P. Roychowdhury, K.-Y. Siu, and A. Orlitsky (editors), pp. 243–293, Kluwer, Boston, 1994.

[10] Valiant, L. G.: A Theory of the learnable. *Communication of the ACM*, **27**:1134–1142, 1984.

[11] Vapnik, V. N., Chervonenkis, A. Y.: On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Application*, **16**:264–280, 1971.