Vapnik-Chervonenkis Dimension of Neural Nets

Wolfgang Maass

Institute for Theoretical Computer Science
Technische Universitaet Graz
Klosterwiesgasse 32/2
A-8010 Graz, Austria
e-mail: maass@igi.tu-graz.ac.at

1 Introduction

Let $\mathcal{N}$ be some arbitrary feedforward neural net with w weights from some weight-space W (e.g. $W = \mathbf{N}, \mathbf{Q}$, or $\mathbf{R}$). If $\mathcal{N}$ has n input-nodes, and if the output gate has range $\{0,1\}$, then $\mathcal{N}$ computes for any weight-assignment $\underline{\alpha} \in W^w$ a function $\mathcal{N}^{\underline{\alpha}}$ from some n-dimensional domain X (e.g. $X = \mathbf{N}^n, \mathbf{Q}^n, \mathbf{R}^n$) into $\{0,1\}$.

One says that a subset S of the domain X is *shattered* by $\mathcal{N}$ if every function $g : S \to \{0,1\}$ can be computed on $\mathcal{N}$, i.e. $\forall\, g : S \to \{0,1\}\ \exists\, \underline{\alpha} \in W^w\ \forall\, x \in S(g(x) = \mathcal{N}^{\underline{\alpha}}(x))$.

The *Vapnik-Chervonenkis dimension* of $\mathcal{N}$ (abbreviated: VC-dimension($\mathcal{N}$); see Cover (1968) for an equivalent definition) is defined as the maximal size of a set $S \subseteq X$ that is shattered by $\mathcal{N}$, i.e.

$$\text{VC-dimension}(\mathcal{N}) := \max\{|S| : S \subseteq X \text{ is shattered by } \mathcal{N}\}.$$

Intuitively one may view the VC-dimension of a neural net $\mathcal{N}$ as the number of "degrees of freedom" that one has in specifying the input/output behaviour of $\mathcal{N}$. Of course one can define without reference to neural nets more generally for *any* class $\mathcal{F}$ of functions $f : X \to \{0,1\}$ (see Vapnik and Chervonenkis, 1971) the VC-dimension of $\mathcal{F}$ by

$$\text{VC-dimension}(\mathcal{F}) := \\ \max\{|S| : S \subseteq X \text{ and } \forall\, g : S \to \{0,1\}\ \exists\, f \in \mathcal{F}\ \forall\, x \in S(g(x) = f(x))\}.$$

Thus our preceding definition of the VC-dimension of a neural net $\mathcal{N}$ is just a special case for the function class $\mathcal{F} := \{f : X \to \{0,1\} : \exists\, \underline{\alpha} \in W^w\ \forall\, x \in X(f(x) = \mathcal{N}^{\underline{\alpha}}(x))\}$. The relevance of the VC-dimension for the training of a neural net can be traced back to the following rather simple mathematical result.

Theorem 1.1 *("Sauer's Lemma", see appendix A2 of Blumer et al., 1989): Let $\mathcal{F}$ be any class of functions from some finite set X into $\{0,1\}$, and set $d :=$ VC-dimension($\mathcal{F}$). Then $\mathcal{F}$ contains at most $\sum_{i=0}^{d} \binom{|X|}{i} \leq |X|^d + 1$ different functions.*

With the help of this result one can establish theoretical results about the *generalization abilities* of a neural net. More precisely one can prove relationships between the "apparent error" of a neural net $\mathcal{N}^{\underline{\alpha}}$ on a randomly drawn training set T, and the "true error" of $\mathcal{N}^{\underline{\alpha}}$ for new examples drawn from the same distribution. This relationship is discussed in PAC-LEARNING AND ARTIFICIAL NEURAL NETWORKS for the idealized setting of the classical PAC-learning model, where one assumes that there exists some assignment $\underline{\alpha}^*$ to the weights of $\mathcal{N}$ such that $\mathcal{N}^{\underline{\alpha}^*}$ has true error 0. We will discuss in the fourth section of this article the corresponding results for the more realistic setting of *agnostic PAC-learning*, where no unrealistic a priori assumption is required.

For either version of the PAC-model one can roughly say that the expected deviation of the *true error* of a trained neural net $\mathcal{N}^{\underline{\alpha}}$ from the *apparent error* of $\mathcal{N}^{\underline{\alpha}}$ on the training set T depends on the size of T *relative* to the VC-dimension of $\mathcal{N}$. Hence it has become of considerable interest to derive estimates for the VC-dimension of various types of neural nets.

We will survey in this article the most important known bounds for the VC-dimension of neural nets that consist of linear threshold gates (section 2) and for the case of neural nets with real-valued activation functions (section 3). In section 4 we discuss a generalization of the VC-dimension for neural nets with non-boolean network-output. With regard to a discussion of the VC-dimension of models for networks of *spiking neurons* we refer to Maass (1994c).

For comparing the asymptotic behaviour of two functions $f, g : \mathbf{N} \rightarrow \mathbf{R}^+$ we use the customary notation $f(n) = O(g(n))$ $[f(n) = \Omega(g(n))]$ if there exists some constant $c > 0$ such that $f(n) \leq c \cdot g(n)$ [resp. $f(n) \geq c \cdot g(n)$] for all sufficiently large n, and $f(n) = \Theta(g(n))$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

2 VC-dimension of Neural Nets with Linear Threshold Gates

A *linear threshold gate* with n inputs computes for given weights $\underline{\alpha} = \langle \alpha_0, \alpha_1, \ldots, \alpha_n \rangle \in W^{n+1}$ the function

$$T^{\underline{\alpha}}(x_1, \ldots, x_n) = \begin{cases} 1 \ , \ \text{if } \sum_{i=1}^{n} \alpha_i x_i + \alpha_0 \geq 0 \\ 0 \ , \ \text{otherwise} \end{cases}$$

from $\mathbf{R}^n$ into $\{0, 1\}$.

Theorem 2.1 *(Wenocur and Dudley): VC-dimension($\mathcal{N}$) $= n + 1$ if $\mathcal{N}$ consists of a single linear threshold gate with n inputs.*

Sketch of the proof: One can easily verify that the set $S := \{\underline{0}\} \cup \{\underline{e}_i : i \in \{1, \ldots, n\}\}$ is shattered by $\mathcal{N}$ (where $\underline{e}_i \in \{0, 1\}^n$ denotes the i-th unit vector). Hence VC-dimension($\mathcal{N}$) $\geq n + 1$.

The lower bound follows from Radon's Theorem, which states that any set S of $\geq n + 2$ points in $\mathbf{R}^n$ can be partitioned into sets S_0 and S_1 such that the convex hulls of S_0 and S_1 intersect. Obviously such sets S_0 and S_1 cannot be separated by any hyperplane, hence not by any threshold gate. ∎

With the help of Theorem 1.1 one can derive from Theorem 2.1:

Corollary: *A linear threshold gate with n inputs can compute at most $|X|^{n+1} + 1$ different functions from any set $X \subseteq \boldsymbol{R}^n$ into $\{0, 1\}$.*

Theorem 2.2 *(Cover, 1968; see also Baum and Haussler, 1989): Let $\mathcal{N}$ be an arbitrary feedforward neural net with w weights that consists of linear threshold gates. Then VC-dimension($\mathcal{N}$)*

$= O(w \cdot \log w)$.

Sketch of the proof: Let S be some arbitrary set of m input-vectors for $\mathcal{N}$. By the Corollary to Theorem 2.1 a gate g in $\mathcal{N}$ can compute at most $|X|^{\text{fan-in}(g)+1} + 1$ different functions from any finite set $X \subseteq \mathbf{R}^{\text{fan-in}(g)}$ into $\{0,1\}$ (fan-in(g) denotes the number of inputs of gate g). Hence $\mathcal{N}$ can compute at most $\prod_{g \text{ gate in } \mathcal{N}} (m^{\text{fan-in}(g)+1} + 1) \leq m^{2w}$ different functions from S into $\{0,1\}$. If S is shattered by $\mathcal{N}$ then $\mathcal{N}$ can compute all 2^m functions from S into $\{0,1\}$. In this case the preceding implies that $2^m \leq m^{2w}$, hence $m \leq 2w \cdot \log m$. It follows that $\log m = O(\log w)$, thus $m = O(w \cdot \log w)$. ∎

It is tempting to conjecture that the VC-dimension of a neural net $\mathcal{N}$ cannot be larger than the total number of weights of all gates in $\mathcal{N}$, which is equal to the sum of the VC-dimensions of the individual gates in $\mathcal{N}$. In view of Theorem 2.1 this conjecture would imply an upper bound $O(w)$ for VC-dimension($\mathcal{N}$) in Theorem 2.2. However the following result (whose proof requires rather complex techniques from circuit theory) shows that the superlinear upper bound of Theorem 2.2 is in fact asymptotically optimal. Hence with regard to the VC-dimension it is fair to say that a neural net can be "more than the sum of its parts".

Theorem 2.3 *(Maass, 1993 and 1994b): Assume that* $(\mathcal{N}_n)_{n \in \mathbf{N}}$ *is any sequence of neural nets with at least 2 hidden layers, where* $\mathcal{N}_n$ *has* n *boolean input nodes and* $O(n)$ *gates. Furthermore assume that* $\mathcal{N}_n$ *has* $\Omega(n)$ *gates on the first hidden layer, and at least* $4 \log n$ *gates on the second hidden layer. We also assume that* $\mathcal{N}_n$ *is fully connected between any two successive layers (hence* $\mathcal{N}_n$ *has* $\Theta(n^2)$ *weights), and that the gates of* $\mathcal{N}_n$ *are linear threshold gates (or gates with the sigmoid activation function* $\sigma(y) = \frac{1}{1+e^{-y}}$*, with round-off at the network output).*
Then VC-dimension$(\mathcal{N}_n) = \Theta(n^2 \cdot \log n)$.

Subsequently Sakurai (1993) has shown that if one allows *real valued* network inputs, then the lower bound of Theorem 2.3 also holds for certain neural nets with *one* hidden layer. In addition he has shown that for the case of real valued inputs one can determine exactly the constant factor in these bounds.

3 VC-dimension of Analog Neural Nets

We consider in this section and the next the case where some gates in $\mathcal{N}$ employ activation functions f with non-boolean output, such as $\sigma(y) = \frac{1}{1+e^{-y}}$ (a gate of fan-in m with activation function f and weights $\alpha_0, \ldots, \alpha_m$ computes the function $\langle y_1, \ldots, y_m \rangle \mapsto f(\sum_{i=1}^{m} \alpha_i y_i + \alpha_0))$. We first consider the case where the network-output of $\mathcal{N}$ is nevertheless boolean-valued (e.g. because the output gate of $\mathcal{N}$ is a linear threshold gate).

It turns out that in order to get upper bounds for the VC-dimension of such neural nets it does not suffice to assume that the analog activation functions in $\mathcal{N}$ are "very smooth squashing functions". Sontag (1992) has shown that for the real-analytic function $\Psi(y) := \frac{1}{\pi}\arctan(y) + \frac{\cos y}{7(1+y^2)} + \frac{1}{2}$ a neural net with 2 real valued inputs, 2 hidden units with activation function Ψ and a linear threshold gate as output gate has *infinite* VC-dimension. Note that this function Ψ is strictly increasing and has limits $1, 0$ at $\pm\infty$ (hence it is a "squashing function"). For the case of neural nets with n *boolean* inputs Sontag constructed activation functions with the same analytic properties as the function Ψ, such that the neural net with the same architecture as above has the maximal possible VC-dimension 2^n.

Thus one cannot hope to prove significant upper bounds for the VC-dimension of an analog neural net if one only knows that its non-boolean activation functions are very smooth strictly increasing squashing functions. More subtle mathematical properties of the activation functions turn out to be crucial, such as the maximal possible number of zeros of any function that is definable from these activation functions with a specific set of operations.

The first upper bound for the VC-dimension of a neural net whose gates employ the activation function $\sigma(y) = \frac{1}{1+e^{-y}}$ is due to Macintyre and Sontag, 1993. By using a sophisticated result from mathematical logic (order-minimality of the elementary theory L of real numbers with the basic algebraic operations and exponentiation) they have shown that the VC-dimension of any finite feedforward neural net with this activation function is finite. Very recently, Karpinski and Macintyre have applied very complicated techniques from differential topology in order to achieve the following upper bound, that is *polynomial* in the number w of weights:

Theorem 3.1 *(Karpinski and Macintyre, 1994): The VC-dimension of any feedforward neural net with the sigmoid activation function σ is bounded by $O(w^4)$, where w is the total number of weights in the neural net. The same upper bound holds for a large class of activation functions that satisfy a certain Pfaffian differential equation.*

For the case of neural nets of arbitrary constant depth with n boolean inputs and polynomially in n many gates with piecewise polynomial activation functions and arbitrary real weights it was shown in (Maass, 1993) that such circuits can be simulated by polynomial size neural nets that consist entirely of linear threshold gates. Hence a polynomial upper bound for the VC-dimension of such neural nets follows immediately from Theorem 2.2. Subsequently Goldberg and Jerrum have shown that with the help of *Milnor's theorem* from algebraic geometry one can prove directly a polynomial upper bound for arbitrary polynomial size neural nets with piecewise polynomial activation functions (in fact their argument also applies to the case of piecewise rational activation functions).

Theorem 3.2 *(Goldberg and Jerrum, 1993): Let $\mathcal{N}$ be any neural net with piecewise polynomial activation functions (with $O(1)$ pieces each), arbitrary real inputs and weights, and boolean output. Then the VC-dimension of $\mathcal{N}$ is at most $O(w^2)$, where w is the total number of weights in $\mathcal{N}$.*

It is an open problem whether this upper bound can be improved to $O(w \log w)$. The best known *lower bound* for the VC-dimension of an analog neural net with piecewise polynomial activation functions (or the activation function σ) is the same bound $\Omega(w \cdot \log w)$ as for the case of neural nets with linear threshold gates (see Theorem 2.3).

4 Generalization of the VC-dimension for Neural Nets with Real-Valued Output

We consider here the case of an analog neural net $\mathcal{N}$ where the range of the activation function of the output gate is *not* boolean-valued. If for example $\mathcal{N}$ has an output gate whose activation function is $\sigma(y) = \frac{1}{1+e^{-y}}$, then it cannot compute *any* function with range $\{0,1\}$. Thus its VC-dimension (as defined above) would be 0. Consequently one has to consider for such neural nets $\mathcal{N}$ a more general notion of a "dimension" in order to give an upper bound for the number of training examples that are needed to train $\mathcal{N}$. A suitable generalization is provided by the notion of a *pseudo-dimension*.

In order to define the pseudo-dimension of a neural net $\mathcal{N}$ one has to specify a loss function ℓ that is used to measure for any example $\langle x, y\rangle \in X \times Y$ the deviation $\ell(\mathcal{N}^{\underline{\alpha}}(x), y)$ of the prediction $\mathcal{N}^{\underline{\alpha}}(x)$ of the neural net from the target value y. Popular choices for ℓ are $\ell(z,y) = |z-y|$, or $\ell(z,y) = (z-y)^2$. One then considers the class $\mathcal{F}_{\mathcal{N},\ell}$ of all functions from $X \times Y$ into $\mathbf{R}$ of the form $\langle x, y\rangle \mapsto \ell(\mathcal{N}^{\underline{\alpha}}(x), y)$ for some weight-assignment $\underline{\alpha} \in W^w$.

One would like to be able to say that $\mathcal{F}_{\mathcal{N},\ell}$ "shatters" a certain subset S of its domain $X \times Y$. However for that one has to generalize the corresponding definition in section 1, since the functions in $\mathcal{F}_{\mathcal{N},\ell}$ may assume other real values besides 0 or 1. This problem is solved by allowing an arbitrary "threshold" $t(\langle x, y\rangle)$ for each element $\langle x, y\rangle$ of the shattered set $S \subseteq X \times Y$ so that for any $f \in \mathcal{F}_{\mathcal{N},\ell}$ one "rounds off" $f(\langle x, y\rangle)$ to 1 if $f(\langle x, y\rangle) \geq t(\langle x, y\rangle)$, and to 0 otherwise.

Definition: *The pseudo-dimension* $dim_P^{\ell}(\mathcal{N})$ *of* $\mathcal{N}$ *with respect to the loss function* ℓ *is defined as the maximal size of a set* $S \subseteq X \times Y$ *which is shattered by* $\mathcal{N}$ *in the sense that there exists some* $t : S \rightarrow \boldsymbol{R}$ *such that*

$$\forall g : S \rightarrow \{0,1\} \;\; \exists\, f \in \mathcal{F}_{\mathcal{N},\ell} \;\; \forall \langle x, y\rangle \in S \;\; (g(\langle x, y\rangle) = 1 \Leftrightarrow f(\langle x, y\rangle) \geq t(\langle x, y\rangle)).$$

Remark: For the special case of the binary range $Y = \{0,1\}$ and the discrete loss function ℓ_D (where $\ell_D(z,y) = 0$ if $z = y$, and $\ell_D(z,y) = 1$ if $z \neq y$) the pseudo-dimension of a neural net $\mathcal{N}$ coincides with its VC-dimension.

If the size m of a training-set $T = (\langle x_i, y_i\rangle_{i \leq m})$ (which is randomly drawn according to some arbitrary distribution D over $X \times Y$) is relatively large in comparison with the pseudo-dimension of $\mathcal{N}$, then the "apparent error" $\frac{1}{m} \sum_{i=1}^{m} \ell(\mathcal{N}^{\underline{\alpha}}(x_i), y_i)$ of $\mathcal{N}^{\underline{\alpha}}$ is (with high probability) close to the "true error" $E_{\langle x,y\rangle \in D}[\ell(\mathcal{N}^{\underline{\alpha}}(x), y)]$ of $\mathcal{N}^{\underline{\alpha}}$, provided that the range of the values $\ell(\mathcal{N}^{\underline{\alpha}}(x), y)$ is bounded. This relationship is made more precise by the following result.

Theorem 4.1 *(Haussler, 1992): Assume that $\mathcal{F}_{\mathcal{N},\ell}$ is a permissible class of functions from $X \times Y$ into some arbitrary bounded interval $[0, B]$ (the "permissibility" of $\mathcal{F}_{\mathcal{N},\ell}$ is a somewhat technical measurability assumption, which is always satisfied if the weightspace W is countable, e.g. for $W \subseteq \boldsymbol{Q}$). Then for any distribution D over $X \times Y$ and any sample $T = (\langle x_i, y_i\rangle)_{i\leq m}$ of m randomly drawn "training-examples" (which are drawn independently according to distribution D) one has for any given $\varepsilon, \delta > 0$ and any sample-size $m \geq \frac{64B^2}{\varepsilon^2}(2\cdot dim_P^\ell(\mathcal{N}) \ln \frac{16eB}{\varepsilon} + \ln \frac{8}{\delta})$ that with probability $\geq 1-\delta$:*

$$\forall\, \underline{\alpha} \in W^w \left(\left| \frac{1}{m} \sum_{i=1}^{m} \ell(\mathcal{N}^{\underline{\alpha}}(x_i), y_i) - E_{\langle x,y\rangle \in D}[\ell(\mathcal{N}^{\underline{\alpha}}(x), y] \right| \leq \varepsilon \right).$$

For neural nets $\mathcal{N}$ with the sigmoid activation function $\sigma(y) = \frac{1}{1+e^{-y}}$ the only known upper bounds for the pseudo-dimension are given by a corresponding generalization of Theorem 3.1 (Karpinski and Macintyre, 1994), and by the following result:

Theorem 4.2 *(Bartlett and Williamson, 1993): Let $\mathcal{N}$ be a neural net with one hidden layer. Assume that the gates of the hidden layer use the activation function σ (alternatively these gates may compute a radial basis function $\underline{y} = \langle y_1, \ldots, y_m\rangle \mapsto e^{-||\underline{y}-\underline{c}||}$ with "weights" $\underline{c} \in \boldsymbol{R}^m$), and that the output gate outputs a weighted sum of its inputs. Then for discrete inputs from $\{-K, \ldots, K\}^n$ the pseudo-dimension of $\mathcal{N}$ is at most $8w \log_2(11 \cdot wK)$, where w denotes the total number of weights in $\mathcal{N}$.*

The *proof* uses an exponential parameter transformation in order to transform the function that is computed by $\mathcal{N}$ into one that is polynomial in its parameters. One can then apply Milnor's Theorem in a similar fashion to that in Theorem 3.2. ∎

For neural nets $\mathcal{N}$ with piecewise polynomial activation functions one can give the following upper bound (Maass, 1994a); which is shown with the help of Milnor's Theorem in the same way as Theorem 3.2.

Theorem 4.3 *$dim_P^\ell(\mathcal{N}) = O(w^2)$ for arbitrary neural nets $\mathcal{N}$ with w real-valued weights and arbitrary piecewise polynomial activation functions that consist of $O(1)$ pieces of degree $O(1)$.*

It is obvious from Theorem 2.3 that the pseudo-dimension of a neural net $\mathcal{N}$ with w weights can be as large as $\Omega(w \cdot \log w)$. It is an open problem whether the pseudo-dimension of a neural net with piecewise polynomial activation functions that consist of $O(1)$ pieces each (or with any other common activation function such as $\sigma(y) = \frac{1}{1+e^{-y}}$) can be any larger.

5 Discussion

The VC-dimension of a neural net with boolean output measures the "expressiveness" of such a neural net. The related notion of a pseudo-dimension provides a similar tool for the analysis of neural nets with real-valued output. The derivation of bounds for the VC-dimension and the pseudo-dimension of neural nets has turned out to be a rather challenging but quite interesting chapter in the mathematical investigation of neural nets. This work has brought a number of sophisticated mathematical tools into this research area, which have subsequently turned out to be also useful for the solution of a variety of other problems regarding the complexity of computing and learning on neural nets (see Roychowdhury et al., 1994, for an overview of the current state of affairs).

Bounds for the VC-dimension (resp. pseudo-dimension) of a neural net $\mathcal{N}$ provide estimates for the number of random examples that are needed to train $\mathcal{N}$ so that it has good generalization properties (i.e., so that the error of $\mathcal{N}$ on new examples from the same distribution is at most ε, with probability $\geq 1-\delta$). From the point of view of a single application-problem these bounds tend to be too large, since they provide such generalization-guarantees simultaneously for *any* probability distribution on the examples and for *any* training algorithm that minimizes disagreement on the training examples. For some special distributions and specific training algorithms one has achieved tighter bounds with the help of heuristical arguments (replica techniques) from statistical physics.

References

Bartlett, P. L., Williamson, R. C., 1993, The VC-dimension and pseudodimension of two-layer neural networks with discrete inputs, Technical Report, Department of Systems Engineering, Australian National University.

Baum, E. B., Haussler, D., 1989, What size net gives valid generalization?, Neural Computation, 1:151-160.

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M. K., 1989, Learnability and the Vapnik-Chervonenkis dimension, J. of the ACM, 36(4):929-965.

Cover, T. M., 1968, Capacity problems for linear machines, in Pattern Recognition, (L. Kanal ed.), Thompson Book Co., 283-289.

Goldberg, P., Jerrum, M., 1993, Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers, Proc. of the 6th Annual ACM Conference on Computational Learning Theory, (ACM-Press, New York), 361-369.

Haussler, D., 1992, Decision theoretic generalizations of the PAC model for neural nets and other learning applications, Information and Computation, 100:78-150.

Karpinski, M., Macintyre, A.,1994, Polynomial bounds for VC-dimension of sigmoidal neural networks, Research Report 85116-CS, Univ. of Bonn; an extended abstract will appear 1995 in the Proc. of EuroCOLT '95, Lecture Notes in Computer Science, Springer (Berlin).

Maass, W., 1993, Bounds for the computational power and learning complexity of analog neural nets (Extended Abstract), Proc. of the 25th Annual ACM Symposium on the Theory of Computing, (ACM-Press, New York), 335-344.

Maass, W., 1994a, Agnostic PAC-Learning of Functions on Analog Neural Nets, Advances in Neural Information Processing Systems, 6:311-318, Morgan Kaufmann (San Mateo); journal version appears in Neural Computation.

Maass, W., 1994b, Neural nets with superlinear VC-dimension, Neural Computation, 6:875-882.

Maass, W., 1994c, On the computational complexity of networks of spiking neurons, Advances in Neural Information Processing Systems, vol. 7, Morgan Kaufmann (San Mateo), to appear.

Macintyre, M., Sontag, E. D., 1993, Finiteness results for sigmoidal "neural" networks, Proc. of the 25th Annual ACM Symposium on the Theory of Computing, (ACM-Press, New York), 325-334.

*Roychowdhury, V. P., Siu, K. Y., Orlitsky, A., 1994, Theoretical Advances in Neural Computation and Learning, Kluwer Academic Publishers (Boston).

Sakurai, A., Tighter bounds of the VC-dimension of three-layer networks, Proc. of WCNN '93, 3:540-543.

Sontag, E. D., 1992, Feedforward nets for interpolation and classification, J. Comp. Syst. Sci., 45:20-48.

Vapnik, V.N., Chervonenkis, A.Y., 1971, On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Applications, 16:264-280.