On the Sample Complexity for Nonoverlapping Neural Networks*

Michael Schmitt†

Lehrstuhl Mathematik und Informatik
Fakultät für Mathematik
Ruhr-Universität Bochum
D–44780 Bochum
Germany
Phone: ++49 234 700-3209
Fax: ++49 234 7094-465
mschmitt@lmi.ruhr-uni-bochum.de

Abstract

A neural network is said to be nonoverlapping if there is at most one edge outgoing from each node. We investigate the number of examples that a learning algorithm needs when using nonoverlapping neural networks as hypotheses. We derive bounds for this sample complexity in terms of the Vapnik-Chervonenkis dimension. In particular, we consider networks consisting of threshold, sigmoidal and linear gates. We show that the class of nonoverlapping threshold networks and the class of nonoverlapping sigmoidal networks on n inputs both have Vapnik-Chervonenkis dimension $\Omega(n \log n)$. This bound is asymptotically tight for the class of nonoverlapping threshold networks. We also present an upper bound for this class where the constants involved are considerably smaller than in a previous calculation. Finally, we argue that the Vapnik-Chervonenkis dimension of nonoverlapping threshold or sigmoidal networks cannot become larger by allowing the nodes to compute linear functions. This sheds some light on a recent result that exhibited neural networks with quadratic VC dimension.

Keywords: Neural networks, read-once formulas, threshold gates, sigmoidal gates, PAC learning, Vapnik-Chervonenkis dimension

---

*Work supported in part by the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2, No. 27150.

†Part of this work was done while the author was with the Institute for Theoretical Computer Science at the Technische Universität Graz, A-8010 Graz, Austria.

1 Introduction

The sample complexity, that is, the number of examples required for a learning algorithm to create hypotheses that generalize well, is a central issue in machine learning. How the sample complexity depends on the structure and the parameters that define a hypothesis class is a question that is often amenable to theoretical investigations. In this paper we study the sample complexity for hypothesis classes consisting of nonoverlapping neural networks. These are feed-forward networks where each node, except the output node, has exactly one outgoing connection. Since such networks have less degrees of freedom, learning using these hypotheses is expected to be more efficient in terms of sample size and computing time than when using unrestricted neural networks.

The computational complexity of learning using nonoverlapping networks has been extensively studied in the literature. (Angluin et al., 1993), for instance, investigated the existence of efficient algorithms that use queries to learn networks that are known as Boolean trees or read-once formulas. In particular, they showed that read-once formulas can be exactly identified in polynomial time using both equivalence and membership queries. They also proved a negative result stating that neither equivalence nor membership queries alone are sufficient to exactly identify all read-once formulas in polynomial time. Research on the learnability of nonoverlapping networks employing neural gates has been initiated by (Golea et al., 1993). They studied the probably approximately correct (PAC) learnability of so-called μ-Perceptron networks with binary weights: A μ-Perceptron network is a disjunction of threshold gates where each input node is connected to exactly one threshold gate. In particular, they designed algorithms that PAC learn these networks in polynomial time from examples only when these examples are randomly drawn under the uniform distribution. (Golea et al., 1996) generalized these results to tree structures in the form of μ-Perceptron decision lists. General nonoverlapping architectures that employ threshold gates as network nodes were considered by (Hancock et al., 1994). They gave a polynomial-time algorithm that PAC learns any nonoverlapping threshold network from examples and membership queries under an arbitrary unknown distribution of the examples.

In this article we investigate the sample complexity for nonoverlapping neural networks in terms of their Vapnik-Chervonenkis (VC) dimension. It is well known that the VC dimension of a function class gives asymptotically tight bounds on the number of training examples needed for PAC learning this class. For detailed definitions and results for this model of learnability we refer the reader to (Anthony and Biggs, 1992; Blumer et al., 1989; Valiant, 1984). Moreover, these estimates of the sample complexity in terms of the VC dimension hold even for agnostic PAC learning, that is, in the case when the training examples are generated by some arbitrary probability distribution (Haussler, 1992). Furthermore, the VC dimension is known to yield bounds for the complexity of learning in various on-line learning models (Littlestone, 1988; Maass and Turán, 1992).

Results on the VC dimension for neural networks abound; see, for instance, the survey by (Maass, 1995). We briefly mention the most relevant ones for this article. Concerning upper bounds, a feedforward network of threshold gates is known to have VC dimension at most $O(w \log w)$ where w is the number of weights (Baum and Haussler, 1989). Networks using piecewise polynomial functions for their gates have VC dimension $O(w^2)$ (Goldberg and Jerrum, 1995) whereas for sigmoidal networks the bound $O(w^4)$ is known (Karpinski and Macintyre, 1997). With respect to lower bounds threshold networks with VC dimension $\Omega(w \log w)$ have been constructed (Sakurai, 1993; Maass, 1994). Furthermore, (Koiran and Sontag, 1997) have shown that there are neural networks with VC dimension $\Omega(w^2)$. Among these are networks that consist of both threshold and linear gates, and sigmoidal networks.

Bounds on the VC dimension for neural networks are usually given in terms of the number of programmable parameters, that are, most commonly, the weights, of these networks. In contrast to the majority of the results in the literature, however, we are not looking at the VC dimension of a single network with a fixed underlying graph, but of the entire class of nonoverlapping networks employing a specified activation function. This must be taken into account when comparing our results with other ones.

The first result on the sample complexity for nonoverlapping neural networks is due to (Hancock et al., 1994). They showed that the VC dimension of the class of nonoverlapping networks having threshold gates as nodes—they called these networks nonoverlapping Perceptron networks—is $O(n \log n)$ where n is the number of inputs.[1] We take this result as a starting line for our work. After introducing the basic definitions in Section 2 we show in Section 3 that the class of nonoverlapping threshold networks has VC dimension $\Omega(n \log n)$, which is—according to the result of (Hancock et al., 1994)—asymptotically tight. Moreover, we show that this bound remains valid even when all networks are required to have depth two and their output gate computes a disjunction. This lower bound is then easily transferred to nonoverlapping networks with sigmoidal gates. In Section 4 we provide a new calculation for an upper bound that considerably improves the constants of (Hancock et al., 1994). Section 5 is a short note on how to derive the upper bound $O(n^4)$ for the class of nonoverlapping sigmoidal networks. Finally, in Section 6 we show that adding linear gates to nonoverlapping threshold or sigmoidal networks cannot increase their VC dimension. Interestingly, it was this use of linear gates that lead to a quadratic lower bound for sigmoidal neural networks in the work of (Koiran and Sontag, 1997). Consequently, if the lower bound $\Omega(n \log n)$ is not tight for nonoverlapping sigmoidal networks one has to look for new techniques in search for asymptotically better bounds.

[1] Since a nonoverlapping neural network on n inputs has $O(n)$ weights, it is convenient to formulate the bounds in terms of the number of inputs. We follow this convention throughout the paper.

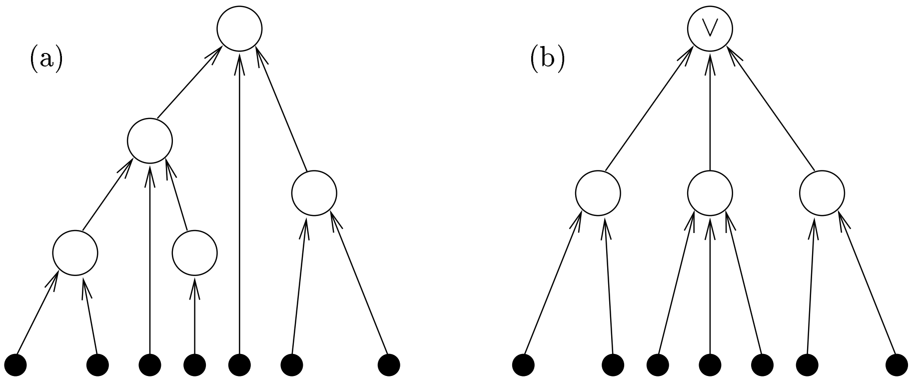

Figure 1: Two examples for nonoverlapping neural networks on seven inputs each. The filled circles are the input nodes, all other ones are computation nodes. Circles with no outgoing connections are output nodes. Network (a) has depth three, network (b) is a depth-two network with a disjunction as output gate such as considered in Section 3. Not shown are the parameters that are associated with each computation node: one weight for each incoming connection and a threshold (except for the output node of network (b) which has no parameters since it is assumed to compute a fixed Boolean disjunction).

2 Basic Definitions

A *nonoverlapping neural network* is a feedforward neural network where there is at most one edge outgoing from each node (see Figure 1). In other words, the connectivity or architecture of the network is a tree.[2] The notion of "nonoverlapping" can be traced back to the work of (Barkai et al., 1990) and has been introduced to model a type of biological neural network where the receptive fields of the neurons do not overlap, i.e., are pairwise disjoint. In a nonoverlapping neural network there is exactly one node, the *output node*, that has no edge outgoing. A *nonoverlapping neural network on n inputs* has n leaves, also called *input nodes*. The *depth* of a nonoverlapping neural network is the length of the longest path from an input node to the output node. The nodes that are not leaves are also known as *computation nodes*. Each computation node has associated with itself a set of $k+1$ real-valued parameters where k is the in-degree of the node: the weights $w_1, \ldots, w_k$ and the threshold t.

We use nonoverlapping neural networks for computations over the reals by assigning functions to its computation nodes and values to their parameters. We consider three types of functions that the nodes may use. All types are obtained

[2] Adopting this notion from graph theory, such a network might most appropriately be called a neural tree. However, this term is avoided here since it is already in use for a completely different classification method based on neural computations along the branches of a decision tree (see, e.g., (Golea and Marchand, 1990; Sirat and Nadal, 1990)).

by applying a so-called activation function to the weighted sum $w_1x_1+\cdots+w_kx_k-t$ where $x_1,\ldots,x_k$ are the input values for the node computed by its predecessors. (The values computed by the input nodes are the input values for the network.) A node becomes a *threshold gate* when it uses as activation function the signum function with $\text{sign}(y)=1$ if $y\geq 0$, and $\text{sign}(y)=0$ otherwise. A *sigmoidal gate* is a node that uses the sigmoidal function $1/(1+e^{-y})$. Finally, a *linear gate* applies the identity function, that is, it simply outputs the weighted sum.

A *nonoverlapping threshold network* is a network where all computation nodes are threshold gates. Correspondingly, we speak of *nonoverlapping sigmoidal networks* and *nonoverlapping linear networks.* If we allow more than one type of activation function for a network then we shall assume that each of its computation nodes may use all types specified. Since we restrict our investigations to networks that compute $\{0,1\}$-valued functions, we assume the output of a network to be thresholded at $1/2$ if the output node is a linear or sigmoidal gate. Thus we can associate with each network on n inputs a set of functions from $\mathbb{R}^n$ to $\{0,1\}$ which are obtained by choosing activation functions for its nodes and varying its parameters over the reals. In a *class of nonoverlapping networks* all members have the same number of inputs, denoted by n, and choose gate functions for their nodes from a specified set, which can be of the three types specified above. The set of functions computed by a class of networks is then defined straightforward as the union of the sets computed by its members.

A *dichotomy* of a set $S\subseteq\mathbb{R}^n$ is a partition of S into two disjoint subsets S_0,S_1 such that $S_0\cup S_1=S$. Given a set $\mathcal{F}$ of functions from $\mathbb{R}^n$ to $\{0,1\}$ and a dichotomy S_0,S_1 of S, we say that $\mathcal{F}$ *induces* the dichotomy S_0,S_1 *on* S if there is a function $f\in\mathcal{F}$ such that $f(S_0)\subseteq\{0\}$ and $f(S_1)\subseteq\{1\}$. We say further that $\mathcal{F}$ *shatters* S if $\mathcal{F}$ induces all dichotomies on S. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{F}$, $\text{VCdim}(\mathcal{F})$, is defined as the largest number m such that there is a set of m elements that is shattered by $\mathcal{F}$.

3 Lower Bounds on the VC Dimension for Nonoverlapping Threshold and Sigmoidal Networks

In this section we consider nonoverlapping neural networks consisting of threshold and sigmoidal gates. We first establish a lower bound on the VC dimension for a class of nonoverlapping threshold networks with certain restrictions: We assume that each network has only two layers of computation nodes and that the output node computes a disjunction (see Figure 1(b) for an example).

Theorem 1 *For each $m,k\geq 1$ there exists a set $S\subseteq\{0,1\}^{m+2^k}$ of cardinality $|S|=m\cdot k$ that is shattered by the class of nonoverlapping threshold networks of depth two that have a disjunction as output gate.*

Proof. Let the set $S \subseteq \{0,1\}^{m+2^k}$ be defined as

$$S = \{e_i : i = 1, \ldots, m\} \times \{d_j : j = 1, \ldots, k\}$$

where $e_i \in \{0,1\}^m$ is the i-th unit vector, i.e., the vector with a 1 in the i-th component and 0s elsewhere, and $d_j \in \{0,1\}^{2^k}$ is specified as follows: Let $A_1, \ldots, A_{2^k}$ be an enumeration of all subsets of $\{1, \ldots, k\}$. We denote the 2^k components of vector d_j by $d_j(1), \ldots, d_j(2^k)$. For each $j \in \{1, \ldots, k\}$ and $l \in \{1, \ldots, 2^k\}$ we define the value of component $d_j(l)$ by

$$d_j(l) = \begin{cases} 1 & \text{if } j \in A_l \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Obviously, S consists of $m \cdot k$ elements. We claim that S is shattered by the class of nonoverlapping threshold networks of depth two having a disjunction as output gate. In order to prove this we show that for each $S' \subseteq S$ there are weights and thresholds for such a network such that this network outputs 1 for elements in S', and 0 for elements in $S \backslash S'$. Fix some arbitrary $S' \subseteq S$. For $i = 1, \ldots, m$ let $\alpha(i)$ be the (unique) element in $\{1, \ldots, 2^k\}$ such that

$$A_{\alpha(i)} = \{j : e_i d_j \in S'\} \ .$$

For convenience we call the input nodes $1, \ldots, m$ the e-inputs, and the input nodes $m+1, \ldots, m+2^k$ the d-inputs. Thus input node $m+l$ becomes d-input l. We employ for each element in the range of α a threshold gate $G_{\alpha(i)}$ that has a connection from d-input $\alpha(i)$ and from none of the other d-inputs. Further, we connect e-input i to gate $G_{\alpha(i)}$ for $i = 1, \ldots, m$. (Notice that this may result in gates that have connections from more than one e-input.) The weights of all connections are fixed to 1 and the thresholds of the gates are set to 2. Obviously, there is at most one connection outgoing from each input node, so that the disjunction of these threshold gates is a nonoverlapping network.

Finally, we verify that the network computes the desired function on S. Suppose that $x \in S'$ where $x = e_i d_j$. The definition of α implies that $j \in A_{\alpha(i)}$. Hence component $\alpha(i)$ of d_j has value $d_j(\alpha(i)) = 1$. Since gate $G_{\alpha(i)}$ receives two 1s — one from e-input i and one from d-input $\alpha(i)$ — the output of the network for $e_i d_j$ is 1.

Assume on the other hand that $e_i d_j \in S \backslash S'$. Then $j \notin A_{\alpha(i)}$ and gate $G_{\alpha(i)}$, which is the only gate that receives a 1 from an e-input, receives only 0s from d-input $\alpha(i)$ (which according to the construction is the only d-input $G_{\alpha(i)}$ is connected to). All other gates G_l, where $l \neq \alpha(i)$, receive at most one 1, which is then from d-input l only. Hence, the output of the network for $e_i d_j$ is 0. $\square$

Choosing $m = n/2$ and $k = \lfloor \log(n/2) \rfloor$ in Theorem 1, we have $m + 2^k \leq n/2 + n/2 = n$. Hence there is a set $S \subseteq \{0,1\}^n$ of cardinality $m \cdot k = \Omega(n \log n)$ that is shattered by the class of networks considered in the above theorem.

Corollary 2 *The VC dimension of the class of nonoverlapping threshold networks on n inputs is $\Omega(n \log n)$. This even holds if all input values are binary and the class is restricted to nonoverlapping threshold networks of depth two with a disjunction as output gate.*

It is well known that in a network that computes a Boolean function a threshold gate can be replaced by a sigmoidal gate without changing the function of the network. (If necessary, the weights have to be scaled appropriately. See, for instance, (Maass et al., 1994) for a treatment in the context of circuit complexity). Thus, the lower bound $\Omega(n \log n)$ also holds for nonoverlapping depth-two networks that may consist of threshold or sigmoidal gates.

Corollary 3 *The class of nonoverlapping depth-two networks on n inputs with threshold or sigmoidal gates has VC dimension $\Omega(n \log n)$. This even holds if the inputs are restricted to binary values.*

We note that depth two is minimal for this lower bound since a threshold gate and a sigmoidal gate both have VC dimension $n+1$: This follows for the threshold gate from a bound on the number of different regions arising from sets of hyperplanes in $\mathbb{R}^n$ which is due to (Schläfli, 1901) (see also (Cover, 1965)). For the sigmoidal gate this follows from the fact that its pseudo dimension is $n+1$ (Haussler, 1992).

Together with the upper bound $O(n \log n)$ due to (Hancock et al., 1994) we obtain asymptotically tight bounds for the class of nonoverlapping threshold networks.

Corollary 4 *The VC dimension of the class of nonoverlapping threshold networks on n inputs is $\Theta(n \log n)$. This holds also for the class of nonoverlapping threshold networks of depth two. Moreover, this bound remains valid for both classes if the inputs are restricted to binary values.*

4 Improved Upper Bounds for Nonoverlapping Threshold Networks

In this section we establish upper bounds for the VC dimension of the class of nonoverlapping threshold networks. Regarding the constants these bounds are better than a previous result derived by (Hancock et al., 1994) which is $13n \log(2en) + 4n \log\log(4n)$.

Theorem 5 *The class of nonoverlapping threshold networks on n inputs, where $n \geq 16e$, has VC dimension at most $6n \log(\sqrt{3}n)$.*

Proof. We estimate the number of dichotomies that are induced by the class of nonoverlapping threshold networks on an arbitrary set of cardinality m. Using the upper bound $(4n)^{n-1}$ on the number of different nonoverlapping networks (or architectures) on n inputs, which was obtained in Lemma 3 of (Hancock et al., 1994) we first derive an upper bound on the number of dichotomies that a single such network induces when all its weights and thresholds are varied.

Let one such network be given and assume without loss of generality that the computation nodes at the lowest level (i.e., those nodes that have input nodes as predecessors) have in-degree 1 and that all other computation nodes have in-degree at least 2. Each of the computation nodes at the lowest level induces at most $2m$ dichotomies on a set of cardinality m. The whole level induces therefore at most $(2m)^n$ different dichotomies. The computation nodes with in-degree at least 2 form a nonoverlapping network that consists of at most $n-1$ nodes and has at most $2n-2$ edges leading to one of these nodes.

According to a result by (Shawe-Taylor, 1995)[3] the number of dichotomies that a threshold network with N computation nodes, partitioned into ν equivalence classes, and W edges induces on a set of cardinality m is at most

$$2^{\nu}\left(\frac{emN}{W-\nu}\right)^{W-\nu} .$$

Using $N = n-1$, $\nu = n-1$, and $W = 2n-2$ we get that the number of dichotomies induced by a nonoverlapping threshold network consisting of $n-1$ computation nodes and $2n-2$ edges is at most $2^{n-1}(em)^{n-1}$.

Putting the bounds together, the total number of dichotomies induced on a set of cardinality m by the class of nonoverlapping threshold networks on n inputs is at most

$$(4n)^{n-1}\cdot(2m)^n\cdot 2^{n-1}(em)^{n-1} .$$

Assume now that a set of cardinality m is shattered. Then

$$\begin{aligned} 2^m &\leq (4n)^{n-1}\cdot(2m)^n\cdot 2^{n-1}(em)^{n-1} \\ &= 2(16en)^{n-1}\cdot m^{2n-1} \\ &\leq 2(mn)^{2n-1} . \end{aligned}$$

For the last inequality we have used the assumption $n \geq 16e$. Taking logarithms on both sides we obtain

$$m \leq (2n-1)\log(mn) + 1 .$$

We weaken this to

$$m \leq 2n\log(mn) . \tag{1}$$

[3] We do not make use of the equivalence relations involved in this result but of the improvement that it achieves compared to (Baum and Haussler, 1989).

Assume without loss of generality that $m \geq \log n$. Then it is easy to see that for each such m there is a real number $r \geq 1$ such that m can be written as $m = r\log(rn)$. Substituting this in (1) yields

$$\begin{aligned} r\log(rn) &\leq 2n(\log(rn\log(rn))) \\ &= 2n(\log(rn) + \log\log(rn)) \qquad (2) \\ &\leq 3n\log(rn) \ . \qquad (3) \end{aligned}$$

The last inequality follows from $\log(rn) \leq \sqrt{rn}$ which holds since $rn \geq 16e$. We divide both sides by $\log(rn)$ and get

$$r \leq 3n \ . \qquad (4)$$

This implies

$$r\log(rn) \leq 3n\log(3n^2) \ .$$

Resubstituting $m = r\log(rn)$ for the left hand side and rearranging the right hand side yields

$$m \leq 6n\log(\sqrt{3}n)$$

as claimed. □

In the statement of Theorem 5 the number of inputs is required to satisfy $n \geq 16e$. We shall show now that we can get the upper bound as close to $4n\log(\sqrt{2}n)$ as we want provided that n is large enough.

Theorem 6 *For each $\varepsilon > 0$, the class of nonoverlapping threshold networks on n inputs has VC dimension at most $4(1+\varepsilon)n\log(\sqrt{2(1+\varepsilon)}n)$ for all sufficiently large n.*

Proof. (Sketch) Fix $\varepsilon > 0$ and $r \geq 1$. For n sufficiently large we have $\log(rn) \leq (rn)^\varepsilon$. Using this in the inequality from (2) to (3) we can infer $r \leq 2(1+\varepsilon)n$ in place of (4). Proceeding similarly as in the last steps of the proof this leads then to the claimed result. □

5 A Note on the Upper Bound for Nonoverlapping Sigmoidal Networks

Using known results on the VC dimension it is straightforward to derive the upper bound $O(n^4)$ for nonoverlapping sigmoidal networks. We give a brief account.

Proposition 7 *The class of nonoverlapping sigmoidal networks on n inputs has VC dimension $O(n^4)$.*

Proof. The VC dimension of a sigmoidal neural network with w weights is $O(w^4)$. This has been shown by (Karpinski and Macintyre, 1997). By Sauer's Lemma (see, e.g., (Anthony and Biggs, 1992)) the number of dichotomies induced by a class of functions with VC dimension $d \geq 2$ on a set of cardinality $m \geq 2$ can be bounded from above by m^d. Thus a nonoverlapping sigmoidal network on n inputs induces at most $m^{O(n^4)}$ dichotomies on such a set. Combining this with the bound $(4n)^{n-1}$ employed in the proof of Theorem 5 and using similar arguments we obtain the bound as claimed. □

6 Nonoverlapping Neural Networks Containing Linear Gates

From (Goldberg and Jerrum, 1995) it has been known that neural networks employing piecewise polynomial activation functions have VC dimension $O(w^2)$, where w is the number of weights. The question whether this bound is tight for such networks has been settled by (Koiran and Sontag, 1997). They have shown that networks consisting of threshold and linear gates can have VC dimension $\Omega(w^2)$. This result was somewhat unexpected since networks consisting of linear gates only compute linear functions and therefore have VC dimension $O(w)$. On the other hand, networks consisting of threshold gates only have VC dimension $O(w \log w)$. This follows from (Cover, 1968) and has also been shown by (Baum and Haussler, 1989). Results that this bound is tight for threshold networks are due to (Sakurai, 1993) and (Maass, 1994).

Therefore, the question arises whether a similar increase of the VC dimension is possible for nonoverlapping threshold or sigmoidal networks by allowing some of the nodes to compute linear functions. We show now that this cannot happen.

Theorem 8 *Let $\mathcal{N}$ be a class of nonoverlapping neural networks consisting of threshold or sigmoidal gates and let $\mathcal{N}^{\text{lin}}$ be a class of neural networks obtained from $\mathcal{N}$ by replacing some of the gates by linear gates. Then* $\text{VCdim}(\mathcal{N}) \geq \text{VCdim}(\mathcal{N}^{\text{lin}})$.

Proof. We show that in a nonoverlapping network of threshold or sigmoidal gates nodes computing linear functions can be replaced or eliminated without changing the function of the network. Assume $\mathcal{N}$ is a nonoverlapping network and y is a node in $\mathcal{N}$ computing a linear function. If y is the output node then y can be replaced by a threshold gate or a sigmoidal gate, where weights and threshold are modified if necessary. (Note that the output of the nonoverlapping neural network is thresholded at $1/2$ for linear and sigmoidal output gates.)

If y is a hidden node (i.e., a computation node that is not the output node) then there is a unique edge e outgoing from y to its successor z. Denote the weight of e by w. Assume that y computes the function $u_1x_1 + \cdots + u_kx_k - t$

where $x_1, \ldots, x_k$ are the predecessors of y, and $u_1, \ldots, u_k, t$ are its weights and threshold. We delete node y and edge e, and introduce k edges from $x_1, \ldots, x_k$ respectively to z. We assign weight wu_i to the edge from x_i for $i = 1, \ldots, n$ and decrease the threshold of z by wt. It can readily be seen that the resulting network is still nonoverlapping and computes the same function as T. $\square$

Combining Theorems 5 and 6 with Theorem 8 we obtain an upper bound for the class of nonoverlapping neural networks with threshold or linear gates.

Corollary 9 *The class of nonoverlapping neural networks on n inputs with threshold or linear gates has VC dimension at most $6n \log(\sqrt{3}n)$ for $n \geq 16e$. Furthermore, for each $\varepsilon > 0$, the VC dimension of this class is at most $4(1 + \varepsilon)n \log(\sqrt{2(1+\varepsilon)}n)$ for all sufficiently large n.*

The technique used in the proof can also be applied to nonoverlapping neural networks that employ a much wider class of gates. If the function computed by a gate can be decomposed into a linear and a non-linear part then the method of deleting a hidden linear node works the same way. Only if the node to be treated is the output node there have to be made some further demands on its function: A sufficient condition is, for instance, monotonicity. If the non-linear part of the gate function is monotonous then a linear output node can be replaced by such a gate without decreasing the VC dimension of the network.

7 Conclusions

Finding methods that incorporate prior knowledge into learning algorithms is an active research area in theoretical and applied machine learning. In the case of neural learning algorithms such knowledge might be reflected in a restricted connectivity of the network generated by the algorithm. We have studied neural networks where each node has at most one outgoing connection and have analyzed the impact that this restriction has on the sample complexity for classes of these networks. The results we derived are given in terms of bounds for their VC dimension.

In this article we have established the asymptotically tight bound $\Omega(n \log n)$ for the class of nonoverlapping threshold networks. We have also derived an improved upper bound for this class. Further, we have considered the implications of having linear gates in nonoverlapping networks: Due to our result demonstrating that the use of linear gates in nonoverlapping threshold networks cannot increase their VC dimension, a known technique to construct networks with quadratic VC dimension does not work for nonoverlapping networks. As a consequence of this, the gap between the currently best known lower and upper bounds for the class of nonoverlapping sigmoidal networks, which are $\Omega(n \log n)$ and $O(n^4)$, is larger than it is for general, i.e. possibly overlapping, sigmoidal networks. To reduce the gap and to extend these investigations to other frequently used types of gates are challenging open problems for future research.

References

Angluin, D., Hellerstein, L., and Karpinski, M. (1993). Learning read-once formulas with queries. *Journal of the Association for Computing Machinery*, 40:185–210.

Anthony, M. and Biggs, N. (1992). *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge.

Barkai, E., Hansel, D., and Kanter, I. (1990). Statistical mechanics of a multilayered neural network. *Physical Review Letters*, 65(18):2312–2315.

Baum, E. B. and Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, 1:151–160.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36:929–965.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334.

Cover, T. M. (1968). Capacity problems for linear machines. In Kanal, L. N., editor, *Pattern Recognition*, pages 283–289, Thompson Book Co., Washington.

Goldberg, P. W. and Jerrum, M. R. (1995). Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148.

Golea, M. and Marchand, M. (1990). A growth algorithm for neural network decision trees. *Europhysics Letters*, 12(3):205–210.

Golea, M., Marchand, M., and Hancock, T. R. (1993). On learning μ-Perceptron networks with binary weights. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems 5*, pages 591–598. Morgan Kaufmann, San Mateo, CA.

Golea, M., Marchand, M., and Hancock, T. R. (1996). On learning μ-Perceptron networks on the uniform distribution. *Neural Networks*, 9:67–82.

Hancock, T. R., Golea, M., and Marchand, M. (1994). Learning nonoverlapping Perceptron networks from examples and membership queries. *Machine Learning*, 16:161–183.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150.

Karpinski, M. and Macintyre, A. (1997). Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176.

Koiran, P. and Sontag, E. D. (1997). Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54:190–198.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318.

Maass, W. (1994). Neural nets with superlinear VC-dimension. *Neural Computation*, 6:877–884.

Maass, W. (1995). Vapnik-Chervonenkis dimension of neural nets. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 1000–1003. MIT Press, Cambridge, MA.

Maass, W., Schnitger, G., and Sontag, E. D. (1994). A comparison of the computational power of sigmoid and Boolean threshold circuits. In Roychowdhury, V., Siu, K.-Y., and Orlitsky, A., editors, *Theoretical Advances in Neural Computation and Learning*, pages 127–151. Kluwer, Boston.

Maass, W. and Turán, G. (1992). Lower bound methods and separation results for on-line learning models. *Machine Learning*, 9:107–145.

Sakurai, A. (1993). Tighter bounds of the VC-dimension of three-layer networks. In *Proceedings of the World Congress on Neural Networks WCNN'93*, volume 3, pages 540–543.

Schläfli, L. (1901). *Theorie der vielfachen Kontinuität*. Zürcher & Furrer, Zürich. Reprinted in: Schläfli, L. (1950). *Gesammelte Mathematische Abhandlungen*. Band I. Birkhäuser, Basel.

Shawe-Taylor, J. (1995). Sample sizes for threshold networks with equivalences. *Information and Computation*, 118:65–72.

Sirat, J. A. and Nadal, J.-P. (1990). Neural trees: a new tool for classification. *Network: Computation in Neural Systems*, 1:423–438.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.