Sample Size Requirements of Feedforward Neural Network Pattern Classifiers*

Terrence L. Fine and Michael J. Turmon
E&TC 388
Cornell University School of Electrical Engineering
Ithaca, NY 14853

We investigate the tradeoffs among network complexity, training set size, and statistical performance of feedforward neural networks.

Nets, labeled as functions $\eta : R^d \to \{0, 1\}$, classify input points $\underline{x} \in R^d$ as either type 0 or type 1. The architecture of all nets under consideration is $\mathcal{N}$, whose size is gauged by its VC dimension v, the size of the largest set of points the architecture can classify in any desired way. Nets $\eta \in \mathcal{N}$ are chosen on the basis of a training set $\mathcal{T} = \{(\underline{x}_i, t_i)\}_{i=1}^n$. These n samples are i.i.d. according to an unknown probability law P. Performance of a network is measured by the error probability

$$\mathcal{E}(\eta) = P(\eta(\underline{x}) \neq t),$$

and a good (perhaps not unique) net in the architecture is

$$\eta^0 = \arg\min_{\eta \in \mathcal{N}} \mathcal{E}(\eta).$$

To select a net using the training set we employ the empirical error frequency

$$\nu_{\mathcal{T}}(\eta) = \frac{1}{n} \sum_{i=1}^{n} |\eta(\underline{x}_i) - t_i|$$

sustained by η on the training set $\mathcal{T}$. A good choice for a classifier is then

$$\eta^* = \arg\min_{\eta \in \mathcal{N}} \nu_{\mathcal{T}}(\eta).$$

By definition $\mathcal{E}(\eta^*) \geq \mathcal{E}(\eta^0)$, and in fact arguments in Vapnik [5] can be adapted to yield the VC upper bound

$$P(\mathcal{E}(\eta^*) - \mathcal{E}(\eta^0) \geq \epsilon) \leq 6 \frac{(2n)^v}{v!} e^{-n\epsilon^2/8}.$$

This inequality shows that sample sizes of about

$$n_c = \frac{16v}{\epsilon^2} \log(\frac{6}{\epsilon})$$

are sufficient to obtain a small probability of a discrepancy of more than ϵ between $\mathcal{E}(\eta^*)$ and $\mathcal{E}(\eta^0)$. If for purposes of illustration we take $\epsilon = .1, v = 50$, we find that $n_c = 328\,000$, which disagrees by orders of magnitude with the experience of practitioners who train such low-complexity networks (about 50 connections).

One way to close this gap between theoretical guidelines and practical experience is to obtain a tighter upper bound. One source of the discrepancy is the union bound employed in the VC development, a tighter version of which is given by Naiman and Wynn [3]:

$$\sum_{1 \leq i \leq N} P(A_i) - \sum_{1 \leq i < j \leq N} P(A_i \cap A_j) \leq P(\bigcup_{1 \leq i \leq N} A_i) \leq$$

$$\sum_{1 \leq i \leq N} P(A_i) - \sum_{1 < i \leq N} P(A_i \cap A_{i-1}).$$

* Prepared with partial support of DARPA under grant number AFOSR-90-0016A.

However, we have shown that these pairwise corrections reduce the upper bound by at most a multiplicative factor of n, which is insignificant compared to other factors entering exponentially, while the lower bound becomes trivial.

The number n_c obtained via VC theory represents a sufficient condition on sample size to obtain reliable classification. To supplement this we have obtained a lower bound or a necessary condition on the training set size needed to obtain reliable classification by examining in detail the error terms for a perceptron under multivariate normal input. Suppose the observed data $\underline{x}$ has equal prior probability of being $N(\mu_0, I_d)$ or $N(\mu_1, I_d)$, and that $n/2$ correctly classified samples are gathered from each prior. When the means are known, the classifier η^0 minimizing error probability is

$$\eta^0(\underline{x}) = 1/2 - \text{sgn}((\underline{x} - (\mu_0 + \mu_1)/2)^T(\mu_0 - \mu_1))/2,$$

and $\mathcal{E}(\eta^0) = \Phi(-\Delta/2)$ where $\Delta^2 = (\mu_0 - \mu_1)^T(\mu_0 - \mu_1)$ and Φ is the distribution of $N(0,1)$. The empirically chosen classifier when the means are unknown is formed by substituting the sample means under each hypothesis, $\bar{x}_0$ and $\bar{x}_1$, into η^0:

$$\eta^*(\underline{x}) = 1/2 - \text{sgn}((\underline{x} - (\bar{x}_0 + \bar{x}_1)/2)^T(\bar{x}_0 - \bar{x}_1))/2.$$

$\mathcal{E}(\eta^*)$ is hard to find (see [1], sec. 6.6), but it can be approximated using arguments in the spirit of Raudys [4]. The condition necessary for reliable classification becomes

$$\mathcal{E}(\eta^*) - \mathcal{E}(\eta^0) \approx \Phi(-(\Delta/2)(1 + 4d/n\Delta^2)^{-1/2}) - \Phi(-\Delta/2) < \epsilon,$$

uniformly over all values of Δ. Analysis reveals that meeting the above condition requires

$$n \geq \frac{v}{33\epsilon^2} \ ,$$

lower than the VC sufficient condition by a factor of order just $\log(1/\epsilon)$. For this special case, it also improves on the necessary condition $n > v/32\epsilon$ obtained by Baum and Haussler [6]. This result confirms that the VC bound is relatively tight, and demonstrates that practitioners are overly optimistic when using small sample sizes.

References

[1] Anderson, T., *An Introduction to Multivariate Statistical Analysis*, second ed., New York: Wiley, 1984.

[2] Baum, E. and D. Haussler, "What size net gives valid generalization?," in D. S. Touretzky, ed., *Advances in Neural Information Processing Systems 1*, 81–90, 1989.

[3] Naiman, D. and Wynn, H., "Inclusion–exclusion–Bonferroni identities...", *Annals of Statistics*, **20**, 43–76.

[4] Raudys, Sh., "On the amount of a priori information in construction of a classification algorithm," *Engineering Cybernetics*, no. 4, 1972. (Russian trans.)

[5] Vapnik, V., *Estimation of Dependences Based on Empirical Data*, New York: Springer, 1982.