Generalization of a Parametric Learning Rule*

Samy Bengio Yoshua Bengio Jocelyn Cloutier Jan Gecsei

Université de Montréal, Département IRO
Case Postale 6128, Succ. "A", Montréal, QC, Canada, H3C 3J7
e-mail: bengio@iro.umontreal.ca

Introduction

We proposed in previous work ([1, 2]) a method to find new learning rules for neural networks, considering them as parametric functions and using any standard optimization method (such as genetic algorithms, gradient descent, and simulated annealing) to select the parameters.

A parametric learning rule is a function of the form $f(x_1, x_2, ...x_n; \theta)$ where the x_i are local variables which can influence the synaptic efficiency, such as the presynaptic and postsynaptic activities, the synaptic weight and the activity of facilitatory (or modulatory) neurons, and θ is a set of parameters, which can be optimized using a cost measuring the performance of the rule over a test set. The question addressed in the next section is whether or not we can find a rule that will be able to learn tasks not used to select the parameters θ.

Capacity of a Parametric Learning Rule

In order for a learning rule obtained through optimization to be useful, it must be successfully applicable in training networks for new tasks. This property of a learning rule is a form of *generalization*, and it can be described using the same formalism used to derive the generalization property of learning systems, based on the notion of *capacity* (see [3] for an introduction).

The capacity h of a parametric learning rule $F(\theta)$ can be intuitively seen as a measure of the cardinality of the set of learning rules it can approximate. It is related to the average generalization ϵ of the rule over new tasks, and the number of training tasks N in the following way. For a fixed number of tasks N, starting from $h = 0$ and increasing it, one finds generalization ϵ to improve (decrease) until a critical value of the capacity is reached. After this point, increasing h makes generalization deteriorate (ϵ increases). For a fixed capacity, increasing the number of training tasks N improves generalization (ϵ asymptotes to a value that depends on h).

We performed experiments on classification problems to study the variation of N, h and the complexity of the tasks, over the learning rule's generalization property (ϵ). We concluded the following from these experiments:

1. The rules found generalized better over similar or simpler tasks (as expected from theory).
2. The generalization (ϵ) improves with the number of tasks (N) used to optimize the rule.
3. For a particular set of tasks, a rule constrained to 7 parameters using a-priori knowledge was better overall than a rule with 16 parameters, in terms of generalization performance and optimization time (as predicted, when we increase the capacity, the generalization can decrease).

References

[1] S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei, *Aspects théoriques de l'optimisation d'une règle d'apprentissage*, in Actes de la conférence Neuro-Nimes 1992, Nimes, France, 1992.

[2] Y. Bengio and S. Bengio, *Learning a synaptic learning rule*, Tech. Rep. 751, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, QC, CANADA, 1990.

[3] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New-York, NY, USA, 1982.

*A longer version of this paper is available by ftp in the neuroprose archive, file bengio.general.ps.Z.

S. Gielen et al. (eds.), *ICANN '93*