Learning Processes of Layered Neural Networks

Sci. Rep. RITU A40 (1995) pp. 313-316
No. 2, January 1995

Learning Processes of Layered Neural Networks

Sumiyoshi FUJIKI and Nahomi M. FUJIKI†
Graduate School of Information Science,
Tohoku University, Sendai 980-77
†*Sendai National Collage of Technology, Sendai 980*

(Received November 11, 1994)

A positive reinforcement type learning algorithm is formulated for a stochastic feed-forward neural network, and a learning equation similar to that of the Boltzmann machine algorithm is obtained. By applying a mean field approximation to the same stochastic feed-forward neural network, a deterministic analog feed-forward network is obtained and the back-propagation learning rule is re-derived.
KEYWORDS: a learning process, feed-forward neural network, stochastic multi-layered network.

1. Introduction

Since Rosenblatt[1] proposed a learning machine, named perceptron, one of the problem has been to find an efficient learning algorithm for a multi-layered network which is capable of solving complicated problems. The error back-propagation[2] algorithm in the multi-layered feed-forward neural network, and the Boltzmann machine algorithm[3] in the symmetrically connected neural network[4] are the basic and typical learning algorithms and architectures of neurocomputers of these days. However, the relation of these models are not revealed sufficiently. The reasons of it would be that the both models were proposed in rather different fields, the models look completely different at a glance in the propagation of neuron activities, and that the both models usually adopt different type neuron units. A unit of the multi-layered feed-forward neural network is required to take a continuous value in order to be differentiable in the back-propagation learning algorithm. On the other hand, a unit in the symmetrically connected neural network usually takes one of two discrete values. Thus the models are different in two points: the synaptic connection, and the value of units. A typical Hopfield model or Boltzmann machine is a symmetrically connected digital unit model, and a multi-layered network adopted in a standard back-propagation algorithm is a feed-forward analog unit model (see Table I). Other choices such as feed-forward digital unit network (considered in the present paper), or symmetrically connected analog unit network may reveal the relation of the two models: the feed-forward analog model and symmetrically connected digital model.

In this paper, we formulate a learning algorithm of a stochastic feed-forward (SFF) digital network in the association problem, and compare it to the Boltzmann machine algorithm. The obtained learning rule is a kind of positive reinforcement with Hebbian and anti-Hebbian learning terms, which are very similar to those of the Boltzmann machine algorithm. By applying a mean field approximation to the model, the model becomes a deterministic analog feed-forward network, and the present learning algorithm reduces to the back-propagation algorithm. Thus the SFF digital network bridges the well-known two models.

2. SFF Model and its Learning Algorithm

We consider a feed-forward neural network consisting of L layers. The lth layer contains $M(l)$ neurons, and each neuron takes one of the two states with a certain probability. The state of the ith neuron on the lth layer is represented by $\sigma_i^{(l)}$ which takes values of ± 1. The 0th layer is an input layer and the Lth layer is an output layer. The neurons on the $l-1$th layer are connected to neurons on the lth layer via synaptic interactions. A state of neurons on the lth layer is represented by $\{\sigma_i^{(l)}\}$. When neurons on the $l-1$th layer are at a state $\{\sigma_i^{(l-1)}\}$, the neurons on the lth layer take a state $\{\sigma_i^{(l)}\}$ with a conditional probability,

$$P(\{\sigma_j^{(l)}\}|\{\sigma_i^{(l-1)}\}) = \prod_j^{M(l)} \left[\frac{e^{\beta\sigma_j^{(l)}h_j^{(l)}}}{2\cosh\beta h_j^{(l)}}\right], \qquad (1)$$

where β is the inverse temperature of the system, and $h_j^{(l)}$ is an internal field at the jth neuron on the lth layer which is given by

$$h_j^{(l)} = \sum_i w_{ji}^{(l)}\sigma_i^{(l-1)} + \theta_j^{(l)}. \qquad (2)$$

Here $w_{ji}^{(l)}$ is the synaptic efficacy from $\sigma_i^{(l-1)}$ to $\sigma_j^{(l)}$, and $-\theta_j^{(l)}$ is the threshold of the neuron $\sigma_j^{(l)}$.

The conditional probability of an output layer to take the state $\{\sigma_k^{(L)}\}$ when an input state is $\{\sigma_i^{(0)}\}$ is given by

$$P(\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\}) = \sum_{\{\sigma^{(1)},\cdots,(L-1)\}} P(\{\sigma_j^{(1)}\},\cdots,\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\}), \qquad (3)$$

where

Table 1: Classification of typical neural network models (and its learning algorithms) by types of a synaptic connection and of a unit.

unit	synaptic connection	
	symmetric connection	feed-forward connection
digital unit	Hopfield model[4] (Boltzmann machine)[3]	stochastic feed-forward (SFF) model (present learning algorithm)
analog unit	mean field approximation[6] of Hopfield model (mean field annealing[7])	deterministic multi-layered network (back-propagation)[2]

$$P(\{\sigma_j^{(1)}\},\cdots,\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\})$$
$$=\prod_j[\frac{e^{\beta\sigma_j^{(1)}h_j^{(1)}}}{2\cosh\beta h_j^{(1)}}]\cdots\prod_k[\frac{e^{\beta\sigma_k^{(L)}h_k^{(L)}}}{2\cosh\beta h_k^{(L)}}]. \quad (4)$$

We consider an association problem where the target patterns are given with probabilities. We assume that the μth pattern is given on the input layer with the probability $p(\mu)$, where $p(\mu)$ satisfies $p(\mu)\geq 0$, and $\sum_{\mu=1}^{p}p(\mu)=1$, and the probability of the νth target pattern on the output layer for the μth input pattern is given by $Q(\{\sigma_k^{(L)}\}_\nu|\{\sigma_i^{(0)}\}_\mu)$. Hereafter in this section and in the next section, we abbreviate conditional probabilities, $Q(\nu|\mu)=Q(\{\sigma_k^{(L)}\}_\nu|\{\sigma_i^{(0)}\}_\mu)$, and $P(\nu|\mu)=P(\{\sigma_k^{(L)}\}_\nu|\{\sigma_i^{(0)}\}_\mu)$, for the simplicity. The quantity which should be minimized by the learning process is a relative entropy averaged over input patterns,

$$\overline{S} \equiv \sum_\mu p(\mu)S_{\rm rel}(\mu),$$
$$S_{\rm rel}(\mu) \equiv -\sum_\nu Q(\nu|\mu)\ln\frac{P(\nu|\mu)}{Q(\nu|\mu)}. \quad (5)$$

The derivatives of the relative entropy with respect to $w_{ji}^{(l)}$ and $\theta_j^{(l)}$ are given by

$$\frac{\delta\overline{S}}{\delta(\beta w_{ji}^{(l)})}=-\sum_\mu p(\mu)\sum_\nu Q(\nu|\mu)$$
$$[<\sigma_i^{(l-1)}\sigma_j^{(l)}>_{\mu\nu}-<\sigma_i^{(l-1)}\tanh\beta h_j^{(l)}>_{\mu\nu}], \quad (6)$$
$$\frac{\delta\overline{S}}{\delta(\beta\theta_j^{(l)})}=-\sum_\mu p(\mu)\sum_\nu Q(\nu|\mu)$$
$$[<\sigma_j^{(l)}>_{\mu\nu}-<\tanh\beta h_j^{(l)}>_{\mu\nu}], \quad (7)$$

where $<O>_{\mu\nu}$ is a weighted average by the conditional probability $P(\{\sigma_j^{(1)}\},\cdots,\{\sigma_k^{(L)}\}_\nu|\{\sigma_i^{(0)}\}_\mu)$.

The decrease of the relative entropy in the learning process is realized by the following change of the weights,

$$\Delta w_{ji}^{(l)}=-\eta\frac{\delta\overline{S}}{\delta(\beta w_{ji}^{(l)})},\quad \Delta\theta_j^{(l)}=-\eta\frac{\delta\overline{S}}{\delta(\beta\theta_j^{(l)})}, \quad (8)$$

for the sufficiently small and positive η, the learning coefficient. The learning process in eqs. (6) and (7) consist of Hebbian like term (first terms) and anti-Hebbian term (second terms). On the averages in eqs. (6) and (7), spin states are counted with the probability of that the desired final pattern is realized, so that the learning is a kind of positive reinforcement.[5]

3. Reformulation of the Boltzmann machine learning algorithm

For the system with the same layered structure but neurons between nearest layers interact symmetrically, we have a similar learning algorithm in the Boltzmann machine. The Boltzmann machine learning algorithm for the association problem was originally formulated by the gradient decent of the relative entropic measure.[3] This algorithm can be reformulated by the minimization of the difference of the free energies of two systems[5], one is the clumped system where neurons on both the input and output layers are clumped in a given pattern μ, the other is a free-end system where only the input layer is fixed and the output layer is kept free. The conditional probability in the free-end system is given by,

$$P(\nu|\mu) = \frac{\sum_{\{\sigma^{(1)},\cdots,(L-1)\}} B(\{\sigma_i^{(0)}\}_\mu, \{\sigma_j^{(1)}\}, \cdots, \{\sigma_k^{(L)}\}_\nu)}{\sum_{\{\sigma^{(1)},\cdots,(L)\}} B(\{\sigma_i^{(0)}\}_\mu, \{\sigma_j^{(1)}\}, \cdots, \{\sigma_k^{(L)}\})}, \quad (9)$$

where B is a Boltzmann factor, $\exp(-\beta E)$, of the system, whose energy E is given by

$$E = -\sum_{l=1}^{L}\sum_j \sigma_j^{(l)}(\sum_i w_{ji}^{(l)}\sigma_i^{(l-1)} + \theta_j^{(l)}). \quad (10)$$

The numerator of eq.(9) is the partition function of the clumped system, and the denominator is the partition function of the free-end system. Thus the logarithm of the conditional probability becomes the difference of the free energies of the clumped system and of the free-end system,

$$\ln P(\nu|\mu) = -(F_{\mu\nu}^c - F_\mu^{f.e.}). \quad (11)$$

Therefore, the relative entropy without constant terms, $\sum_\nu Q(\nu|\mu)\ln Q(\nu|\mu)$, is equivalent to the difference of free energies averaged over input patterns,

$$\overline{S} \equiv \sum_\mu p(\mu)\sum_\nu Q(\nu|\mu)(F_{\mu\nu}^c - F_\mu^{f.e.}). \quad (12)$$

Where $F_{\mu\nu}^c$ and $F_\mu^{f.e.}$ are free energies of the clumped and free-end systems. The derivatives of $\overline{S}$ with respect to $w_{ji}^{(l)}$ and $\theta_j^{(l)}$ are given by

$$\frac{\delta\overline{S}}{\delta(\beta w_{ji}^{(l)})} = -\sum_\mu p(\mu)\sum_\nu Q(\nu|\mu)$$
$$[< \sigma_i^{(l-1)}\sigma_j^{(l)} >_{\mu\nu}^c - < \sigma_i^{(l-1)}\sigma_j^{(l)} >_\mu^{f.e.}], \quad (13)$$

$$\frac{\delta\overline{S}}{\delta(\beta\theta_j^{(l)})} = -\sum_\mu p(\mu)\sum_\nu Q(\nu|\mu)$$
$$[< \sigma_j^{(l)} >_{\mu\nu}^c - < \sigma_j^{(l)} >_\mu^{f.e.}], \quad (14)$$

where $< O >_{\mu\nu}^c$ and $< O >_\mu^{f.e.}$ are weighted averages by the Boltzmann weights. The learning process to minimize $\overline{S}$ is performed by updating weights and thresholds as follows:

$$\Delta w_{ji}^{(l)} = -\eta\frac{\delta\overline{S}}{\delta(w_{ji}^{(l)})}, \quad \Delta\theta_j^{(l)} = -\eta\frac{\delta\overline{S}}{\delta(\theta_j^{(l)})}. \quad (15)$$

The equations (12 ~ 14) are very similar to eqs. (6 ~ 8), except the calculations of thermal averages: the one is averaged over the conditional probability with which a desired answer is obtained, and the other over the Boltzmann factors. At a glance, the second terms (anti-Hebbian terms) of eqs. (12) and (13) seem different from those of eqs. (6) and (7), but these terms are related physically, since the free-end condition in the symmetrically connected network corresponds to the free-end probabilities (including all the right and wrong answers on the output layer) in the feed-forward network, and the thermal average of $\sigma_j^{(l)}$ with the free-end probabilities is given by $\tanh\beta h_j^{(l)}$.

4. A mean field approximation to the SFF model

By applying a mean field approximation to each layer successively, we have a deterministic analog feed-forward network where the thermal average of $\sigma_j^{(l)}$, denoted by $m_j^{(l)}$, is determined by thermal averages of $\sigma_i^{(l-1)}$ as

$$m_j^{(l)} = \tanh\{\beta(\sum_i w_{ji}^{(l)} m_i^{(l-1)} + \theta_j^{(l)})\}. \quad (16)$$

In the mean field approximation, the conditional probability $P(\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\}_\mu)$ in eq.(5) is factorized by conditional probabilities $p(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu)$ for each output unit k as $P(\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\}_\mu) = \prod_k p(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu)$. Assuming the same factorization to the desired probability as $Q(\{\sigma_k^{(L)}\}|\{\sigma_i^{(0)}\}_\mu) = \prod_k q(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu)$, eq.(5) is reduced to

$$\overline{S}_{\rm mfa} \equiv -\sum_\mu p(\mu)\sum_k\sum_{\sigma_k^{(L)}} q(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu) \times \ln p(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu). \quad (17)$$

Replacing $p(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu)$ by $m_{k,\mu}^{(L)}$, and $q(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu)$ by $\zeta_{k,\mu}^{(L)}$ by

$$p(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu) = \frac{1}{2}(1+\sigma_k^{(L)}m_{k,\mu}^{(L)}),$$
$$q(\sigma_k^{(L)}|\{\sigma_i^{(0)}\}_\mu) = \frac{1}{2}(1+\sigma_k^{(L)}\zeta_{k,\mu}^{(L)}), \quad (18)$$

derivatives are given by

$$\frac{\delta\overline{S}_{\rm mfa}}{\delta(\beta w_{ji}^{(l)})} = -\sum_\mu p(\mu)\delta_{j,\mu}^{(l)}m_{i,\mu}^{(l-1)},$$
$$\frac{\delta\overline{S}_{\rm mfa}}{\delta(\beta\theta_j^{(l)})} = -\sum_\mu p(\mu)\delta_{j,\mu}^{(l)}, \quad (19)$$

where

$$\delta_{j,\mu}^{(l-1)} \equiv \beta\tanh'(h_{j,\mu}^{(l-1)})\sum_k \delta_{k,\mu}^{(l)}w_{k,j}^{(l)}, \quad \text{for } l = 2, \ldots, L, \quad (20)$$

with $\delta_{j,\mu}^{(L)} = \zeta_{j,\mu}^{(L)} - m_{j,\mu}^{(L)}$, $\tanh' x = 1 - \tanh^2 x$, and $h_{j,\mu}^{(l)} = \sum_i w_{ji}^{(l)}m_{i,\mu}^{(l-1)} + \theta_j^{(l)}$. Thus we have the standard back-propagation learning equation, except only that the derivative of the activation function on the output layer is missing.

The similar equations in this section were given by Hopfield[6] as equations for an analog perceptron, which

is a deterministic analog feed-forward network. Hopfield, however, proposed a quantity $\overline{S}_{\rm mfa}$ rather intuitively, and did not derive these equations as a mean field approximation to the stochastic feed-forward digital network.

5. Conclusions and Discussion

We formulated a learning algorithm of a stochastic feed-forward (SFF) digital neural network. By minimizing an entropic measure which is analogous to the Kullback divergence, a learning rule similar to that of the Boltzmann machine is obtained. The present learning rule consists of Hebbian and anti-Hebbian terms, and the learning is a kind of positive reinforcement.

Numerical calculations of the SFF network on AND and XOR problems show very similar results to those of the Boltzmann machine.[8] This similarity suggests that the learning algorithm of the Boltzmann machine can be interpreted by the positive reinforcement mechanism. In other words, the backward connection of the Boltzmann machine can be explained as an effective feed-back resulting from the positive reinforcement.

The application of the mean field approximation to the present stochastic feed-forward digital network results in a deterministic analog feed-forward network, and the learning rule similar to the standard back-propagation algorithm is obtained. Thus the present model bridges the two well-known neural network models with the learning rules, namely the Boltzmann machine algorithm in the symmetrically connected digital network and the back-propagation algorithm in the feed-forward analog network.

The positive reinforcement learning mechanism for the present model may be explained by a synaptic global feed-back loop. This learning mechanism might be possible by also a chemical feed-back loop such as emitting chemicals around the target units when the desired pattern is realized and the chemicals reinforce the synaptic weights and thresholds which gave a right answer.

We are now proceeding the learning processes of more complex problems in the stochastic feed-forward and Boltzmann machine networks. Preliminary numerical calculations show that 1) the both networks learn similarly even in these complex problems, 2) the both networks are robust: damages of neurons are recovered by other neurons, 3) well deal with the generalization problem, and 4) the optimized construction of the network is achieved automatically: an automatic recruitment of other useful units and a detachment of useless units in the learning process. These are very important features of the real brain function as well as in the design of neurocomputers.

1) F. Rosenblatt: Psychol. Rev. **65** (1958) 386.

2) D. E. Rumelhart, G. E. Hinton and R. J. Williams: in *Parallel Distributed Processing* eds. D. E. Rumelhart and J. L. McClelland (1986) MIT Press.

3) D. H. Ackley, G. E. Hinton and T. Sejnowski: Cognitive Science **9** (1985) 147.

4) J. Hopfield: Proceedings of the National Academy of Sciences, **79** (1982) 2554.

5) J. Hertz, A. Krogh and R. G. Palmer: in *Introduction to the Theory of Neural Computation* (1991) Addison-Wesley Publishing Co.

6) J. Hopfield: Proceedings of the National Academy of Sciences **84** (1987) 8429.

7) C. Peterson and J. R. Anderson: Complex Systems **1** (1987) 995.

8) S. Fujiki and N. M. Fujiki: J. Phys. Soc. Japan **64** No. 3 (1995).