Neural Computing* (Extended Abstract)

Shun-ichi AMARI**

Key Words : Neural computing, Single neuron, Network, Loss function, Learning algorithm, self-organization

1. Introduction

It is important to present mathematical theories of neural computing in a unified manner to be applicable to various architectures of networks (Amari, 1990). The present paper focusses on the learning aspects of neural computing. The theory elucidates the manner of parameter modification driven by a time series of input signals generated from a stochastic information source. A network modifies its behavior such that it adapts to the environmental information structure.

2. Learning of a Single Neuron

We begin with learning of a single neuron. Let us consider a mathematical neuron model which receives n input signals (a vector input) $\mathbf{x}=(x_1, \dots, x_n)$ and emits an output signal z . Its input-output relation is given by

$$z = f(\mathbf{w} \cdot \mathbf{x} - h),$$

where $\mathbf{w}=(w_1, \dots, w_n)$ is the synaptic weight vector, h is a threshold, and f is an output function.

Let us consider an information source emitting signals $\mathbf{x}$ with probability $p(\mathbf{x})$, each time independently. Let $y=y_d(\mathbf{x})$ be the desired output when $\mathbf{x}$ is input. The desired signal is missing in the case of self-organization, and is sometimes stochastic including noises. An information source is denoted by $I=\{\mathbf{x}, y, p(\mathbf{x}, y)\}$, where $p(\mathbf{x}, y)$ is a probability distribution of an input-output pair. A neuron modifies its synaptic weight vector by receiving signals from I to adapt to it.

A general learning equation can be written as

$$\mathbf{w}_{t+1} = (1 - \epsilon) \mathbf{w}_t + \epsilon r_t \mathbf{x}_t,$$

where $\mathbf{w}_t$ is the synaptic weight at time t , $\mathbf{x}_t$ is the input signal at t , ϵ is a small constant, and r_t is called a learning signal, which designates when modification takes place. In most learning algorithms proposed so

far, the learning signal is determined as a function of the form

$$r_t = r(\mathbf{w} \cdot \mathbf{x}, y)$$

It was proved (Amari, 1977) that there exists a potential function

$$l(\mathbf{w}, \mathbf{x}, y) = \frac{1}{2} \epsilon |\mathbf{w}|^2 - \int_0^{\mathbf{w} \cdot \mathbf{x}} r(u, y) du$$

such that the learning equation is written as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \epsilon \partial l(\mathbf{w}_t, \mathbf{x}_t, y_t) / \partial \mathbf{w}.$$

Since input signals are emitted from the information source I the learning behavior is described by the average of the above equation with respect to the distribution $p(\mathbf{x}, y)$. This yields the average learning equation

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \epsilon \partial L(\mathbf{w}_t) / \partial \mathbf{w},$$

where

$$L(\mathbf{w}) = \int l(\mathbf{w}, \mathbf{x}, y) p(\mathbf{x}, y) d\mathbf{x} dy.$$

This is the average learning equation proposed by Amari (1977). The synaptic weight vector converges to one of the local minima of the learning potential function $L(\mathbf{w})$. Various examples, such as Hebbian learning, principal component analyzer, least square learning, were analyzed.

3. Learning of a Network

It is possible to generalize this idea to obtain learning algorithms of neural networks. We first consider a feedforward network. It has a very simple architecture, in which neurons are connected in a feedforward manner without feedback connections. This network receives an input signal $\mathbf{x}$, processes it, and finally outputs a vector signal $\mathbf{z}$. Therefore, it realizes a transformation or a mapping from the input signal space $X=\{\mathbf{x}\}$ to the output signal space $Z=\{\mathbf{z}\}$.

The mapping depends on a number of network parameters such as synaptic weights and thresholds, some of which are modifiable. We denote all the modifiable network parameters by a vector $\boldsymbol{\theta}=(\theta_1, \theta_2, \dots, \theta_N)$, and denote the network mapping by

$$\mathbf{z} = f(\mathbf{x}; \boldsymbol{\theta}).$$

Input signals $\mathbf{x}$ are generated at an information source I and are applied to neurons or neural net-

* Special talk at International Symposium on Mineral Exploration: -The Use of Artificial Intelligence, 1990 (ISME-AI '90, Tokyo and Tsukuba)

** Fac. of Eng., Univ. of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113, Japan

works. Signals are assumed to be randomly generated subject to a probability distribution $p(\mathbf{x})$, where $p(\mathbf{x})$ represents the probability or relative frequency of signal $\mathbf{x}$ in the environment to which a neural network is to adapt by learning.

It is also assumed here that the information source I emits signal $\mathbf{x}_t$ at time t , each time independently subject to a fixed probability distribution $p(\mathbf{x})$. In some cases, a desired output $\mathbf{y} = \mathbf{y}_d(\mathbf{x})$ is provided together with $\mathbf{x}$ to help a network to improve its behavior. Here $\mathbf{y}_d(\mathbf{x})$ is called the desired output signal to input $\mathbf{x}$. This is regarded as a teacher signal.

In some other cases, no teacher signals are given, so that a network should modify its behavior based on input signal $\mathbf{x}$ and the state of the network processing it. Since $\mathbf{x}$ is generated subject to the environmental probability structure $p(\mathbf{x})$ of the signal space, it is still possible for a network to modify its behavior to fit well with the environmental structure $p(\mathbf{x})$. This scheme is said to be unsupervised learning or self-organization (Amari and Takeuchi, 1978).

4. Input-Output Behaviors of Networks

Before describing learning algorithms, some examples of parametric input-output behaviors $f(\mathbf{x}; \boldsymbol{\theta})$ are given here. The first example is a multilayer perceptron. Let z_i^l be the i -th output of the l -th layer, and let $w_{ij}^{(l-1)}$ be the connection weight from the j -th output of the $(l-1)$ th layer to the i -th neuron of the l -th layer. We then have

$$z_i^l = f\left(\sum_j w_{ij}^{(l-1)} z_j^{l-1} - h_i^l\right).$$

Since the input z^0 to the first layer is $\mathbf{x}$ itself and the output from the last layer is the total output, the above equation recursively defines the total input-output relation $\mathbf{z} = f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ consists of all the $w_{ij}^{(l-1)}$ and h_i^l .

The second example is a radial basis function, which might not be a neural network in a narrow sense but our learning theory is applicable to this case. Let $k(\mathbf{x})$ be a given function whose typical example is

$$k(\mathbf{x}) = \exp\left\{-\frac{1}{2}|\mathbf{x}|^2\right\}.$$

When the scalar output z is written as

$$z = \sum w_i k(\mathbf{x} - \mathbf{u}_i),$$

where $\boldsymbol{\theta} = (w_1, \dots, w_n; \mathbf{u}_1, \dots, \mathbf{u}_n)$ is the set of modifiable parameters, the above relation is written as

$$z = f(\mathbf{x}; \boldsymbol{\theta}).$$

It is easy to extend this to the vector output case. This is called the radial basis function expansion. The problem is to obtain the optimal $\boldsymbol{\theta}$ from examples.

The third example is a pattern recognition system. We consider the following problem of pattern recognition: Let $C_\alpha, \alpha = 1, \dots, m$, be m categories, to which signals $\mathbf{x}$ belong. Let $p_\alpha(\mathbf{x})$ be the probability density function of signals belonging to category C_α and let p_α be the prior probability that a signal $\mathbf{x}$ belongs to C_α . The posterior probability $p(\alpha|\mathbf{x})$ that a given $\mathbf{x}$ belongs to C_α is given by

$$p(\alpha|\mathbf{x}) = p_\alpha p_\alpha(\mathbf{x}) / \sum_\beta p_\beta p_\beta(\mathbf{x}).$$

It is well known that the best decision (in the sense of the least misclassification) is to assign $\mathbf{x}$ to the category C_β that maximizes $p(\alpha|\mathbf{x})$, i. e., such C_β that

$$\max_\alpha p(\alpha|\mathbf{x}) = p(\beta|\mathbf{x})$$

Now we consider a layered neural network, which has m output neurons corresponding to the categories. Let $z_\beta(\mathbf{x})$ be the output of the β -th neuron corresponding to the β -th category C_β . The network decides that $\mathbf{x}$ belongs to C_β , when

$$\max_\alpha |z_\alpha(\mathbf{x})| = z_\beta(\mathbf{x}).$$

Let us put

$$q(\alpha|\mathbf{x}) = \exp|z_\alpha(\mathbf{x}) - z(\mathbf{x})|,$$

where $z(\mathbf{x})$ is the normalizing constant such that

$$\sum_\alpha q(\alpha|\mathbf{x}) = 1.$$

We regard that the outputs $f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{z}_\beta(\mathbf{x}, \boldsymbol{\theta})$ of this network calculate the logarithm $\log p(\beta|\mathbf{x})$ of the desired probabilities, $q(\alpha|\mathbf{x})$ approximating the true $p(\alpha|\mathbf{x})$.

5. General Loss Functions

Returning to the general theory, let $l(\mathbf{x}; \boldsymbol{\theta})$ be a loss function, when signal $\mathbf{x}$ is processed by a network whose parameter is $\boldsymbol{\theta}$. When a teacher signal $\mathbf{y}$ exists, it depends on $\mathbf{y}$, too, so that it is written as $l(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$. A simple example of the loss is the squared error,

$$l(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \frac{1}{2} |f(\mathbf{x}; \boldsymbol{\theta}) - \mathbf{y}|^2,$$

where $\mathbf{y}$ designates the desired output

$$\mathbf{y} = \mathbf{y}_d(\mathbf{x}).$$

However, this is not the only possibility and we can use many other types of loss. In the case of a pattern recognition network, since the *Kullback-Leibler* measure gives a natural divergence measure between two probability distributions, the loss to be minimized is

$$\begin{aligned} D(\mathbf{p}, \mathbf{q}; \mathbf{x}) &= \sum_\beta p(\beta|\mathbf{x}) \log \frac{p(\beta|\mathbf{x})}{q(\beta|\mathbf{x})} \\ &= \sum_\beta p(\beta|\mathbf{x}) [\log p(\beta|\mathbf{x}) - z_\beta(\mathbf{x}) + z(\mathbf{x})]. \end{aligned}$$

Therefore ,

$$l(\mathbf{x}, \beta; \boldsymbol{\theta}) = -z_{\beta}(\mathbf{x}; \boldsymbol{\theta}) + z(\mathbf{x}; \boldsymbol{\theta})$$

gives a loss function when signal $\mathbf{x}$ belonging to C_{β} is processed by a network whose parameter is specified by $\boldsymbol{\theta}$, where the teacher signal specifies the true category C_{β} .

The overall behavior of a network is evaluated by the expected loss

$$L(\boldsymbol{\theta}) = \langle l(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \rangle,$$

where $\langle \rangle$ denotes the expectation with respect to the probability distribution $p(\mathbf{x}, \mathbf{y})$ of the environmental information source.

6. Learning Algorithm

In order to obtain the optimal network by learning, a general learning scheme is proposed by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - c_t \partial l(\mathbf{x}_t, \mathbf{y}_t; \boldsymbol{\theta}_t) / \partial \boldsymbol{\theta},$$

where c_t is scalar (or a positive-definite matrix), and the suffix t denotes the time. By this algorithm, $\boldsymbol{\theta}_t$ approaches one of the local minima of $L(\boldsymbol{\theta})$, so that the global optimization may not be attained. If c_t is chosen properly as the stochastic approximation technique suggests, the network converges to one of the local minima in probability. This idea of general learning is proposed by Amari (1967).

If we apply this idea to a layered network of feedforward connections with the squared error loss, we have the famous backpropagation learning algorithm.

The learning algorithm can be written as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - c_t \partial l(\mathbf{x}_t, \beta; \boldsymbol{\theta}_t) / \partial \boldsymbol{\theta}$$

when input signal $\mathbf{x}_t$ at time t belongs to C_{β} . This opens a new learning paradigm other than the backpropagation learning.

7. Self-organization of Neural Networks

Finally, a theory of self-organizing neural field is presented. Given an information source, I , of signals, it is possible that a neural network automatically forms an inner representation of signals by self-organization. In particular, when each signal in I is represented by a local excitation at a specific position of a neural field, we have a map of the signals on the neural field. This is called a cortical map or a neural information map. It is interesting to know the basic properties of a map; the topological relation be-

tween the map and the original signal space X , the resolution of a map, the magnification factor of each signal in the map, etc.

A series of mathematical analysis has been performed based on the above general principle of neural learning. The theory predicts various interesting phenomena:

(1) magnification property. When a signal is applied more frequently, its map becomes larger in the cortical representation.

(2) inverse law. The size of the receptive field of a neuron has an inverse relation with the size of the cortical area which a signal excites.

(3) topology preserving law. Two signals which are close in the signal space are apt to be mapped to close locations in the cortical field.

(4) formation of patch or columnar microstructure. Even when the signal space is a continuum, it can be automatically divided into patches or columns in its cortical map.

We have succeeded in analyzing the above effect mathematically by using a simple model of a self-organizing neural field (Amari, 1980, 1983; Takeuchi and Amari, 1979)

References

- Amari, S. (1967) Theory of adaptive pattern classifiers. *IEEE Trans.*, **EC-16**, 299 - 307.
- Amari, S. (1977) Neural theory of association and concept-formation. *Biological Cybernetics*, **26**, 175 - 185.
- Amari, S. (1980) Topographic organization of nerve fields. *Bull. of Math. Biology*, **42**, 339 - 364.
- Amari, S. (1983) Field theory of self-organizing neural nets. *IEEE Trans. Systems, Man and Cybernetics*, **SMC-13**, 741 - 748.
- Amari, S. (1990) Mathematical foundations of neuro-computing. *Proc. IEEE Special Issue on Neural Networks*, **78**, 1443 - 1463.
- Amari, S. and Takeuchi, A. (1978) Mathematical theory on formation of category detecting nerve cells. *Biological Cybernetics*, **29**, 127 - 136.
- Takeuchi, A. and Amari, S. (1979) Formation of topographic maps and columnar microstructures. *Biological Cybernetics*, **35**, 63 - 72.

(Received 10 January 1991; accepted in revised form 14 May 1991)

要 旨

ニューラル・コンピューティング — 神経回路網の数理 — : 甘利 俊一

ニューラル・コンピューティングの数学的理論を作り、種々のアーキテクチャの回路網の性質を統一的理解することが重要である。ここではニューラル・コンピューティングのうち、学習に焦点をあてる。ある情報源から確率的な時系列として発生する入力信号をもとに、自己の結合のパラメータの値を回路網がどう調整していくかを明らかにする。これにより回路網は外界の情報構造に自己を適合させるのである。まず、1個の神経細胞の学習過程を明らかにする。このために学習ポテンシャル関数を導入し、これにより各入力信号の出力に及ぼす重みを表すベクトルが、学習によりそのポテンシャル関数の極小値に確率的に収束することを示す。次にこの考えを前向き結合の多層神経回路網の学習問題に使えるように一般化する。とくに、教師信号のある場合に、回路網の動作の平均損失関数を一般的に定式化し、これをもとに具体的な学習のアルゴリズムを与える。こうして、確率的な降下法による学習神経回路網の一般的な定式化ができる。損失関数を出力の平均二乗誤差にとることにより、このアルゴリズムからバックプロパゲーション学習を特殊な場合として導くことができる。この他、自己回帰結合のある一般のニューラル・ネットワークの自己組織を教師信号なしの学習過程として定式化できる。これは皮質における情報表現、すなわち信号配置図が学習によってどう定まっていくかを明らかにする。とくに、皮質におけるトポロジーの保持、情報表現の拡大率、コラム構造の自己組織的形成など、多くの興味ある性質が数理的に明かになる。

キーワード：ニューラル・コンピューティング、神経細胞、ネットワーク、損失関数、学習アルゴリズム、自己組織

(1991年1月10日受付, 1991年5月14日受理)