知識工学と人工知能 67 － 2
（１９８９．１１．８）

誤差最小化BAMによるパターン認識

林　幸雄
富士ゼロックスシステム技術研究所

概要

BAM(Bidirectional Associative Memory)モデルをパターン認識に適用する為、従来のBAMの自己組織化学習に認識誤差最小化学習を付加して認識辞書を学習するとともに、文字間のクロストークを低減するように認識辞書である入出力間の結合重みを各文字のクラスタ中心で初期設定する方法を提案する。簡単な文字認識実験の結果、本モデルは従来のBPモデルに対して学習回数と汎化能力の面で優位であった。アトラクターの引き込みによりパターンを識別する本モデルは、新しいタイプのクラスタリング手法として期待できよう。

Pattern Recognition by Least Error Bidirectional Associative Memory

Yukio HAYASHI
System Technology Research Lab. FUJI XEROX Co., LTD.
2274, Hongo, Ebina-shi, Kanagawa 243-04, Japan

ABSTRACT

Two improvements of Bidirectional Associative Memory are proposed in order to apply to character recognition.
First one is the addition on least error learning to minimize recognition error.
Second is the definition of initial connection based on each clustering center.
As a simple result of simulation, this model is superior to conventional BP model in the iteration time of learning and the ability of generalization.
This method, which discriminates a character by pulling into an attractor, is considered as a new recognition approach.

1. Introduction

Human brain has a superior ability for dealing with incomplete information. As a technical realization of this ability, Bidirectional Associative Memory are suggested [1][2]. But this model has both advantages and disadvantage, e.g. flexible data supplement for incomplete information and the pattern discrimination criterion depending on patterns themselves [3][4][5]. On the other hand, as a view of pattern recognition, Back Propagation (BP) model is promising and there are good results for many applications. Funahashi points out the theoretical ability of BP model [6], and Kawahara / Irino expand the idea for "SPAN", which is a setting method of the good initial connection value for a layer model [7]. Katagiri also explains the relation between BP model and Vector Quantization method (VQ) by introducing the metric representation like SPAN. BP model is understood to have the advantage of pattern discrimination criterion by least error, but it is trapped into a bad local minimum in many case, and spend much time for learning as disadvantages.

In this paper, I consider a recognition process as an interactive process between input and output patterns, and propose Least error Bidirectional Associative Memory (L-BAM) model which utilize advantages of both BP model and BAM model complementarity in order to apply BAM model to pattern recognition. The improvement points are the addition of least error learning like BP model to self-organization learning of BAM model, proof of the stability of learning, and the way to set the initial connection value in order to decrease the crosstalk. As a simple result of character recognition, L-BAM is superior to BP model in the iteration time of learning and the data complementary ability (generalization).

L-BAM is considered as a nonlinear discrimination method by pulling into an attractor. This method can be expected as a new type of clustering.

2. L-BAM

In the conventional BAM model, the connection value is adaptively updated by learning, but the output error can not anytime minimized, since the learning method is self-organization learning like Hebbian learning based correlation between input and output. In chapter 2.1, I propose L-BAM model where the least error term is added with teacher signal to conventional BAM model, holding the local processing like Hebbian learning. Next, in chapter 2.2, I prove the stability of an energy function in learning phase, in chapter 2.3, show the meaning of the energy function.

2.1 Proposal of L-BAM

L-BAM is a two layer network of symmetrically interconnected neurons (as shown in Fig.1). There are N neurons in A = $\{a_1,....a_N\}$, P neurons in B = $\{b_1,...,b_P\}$, and the NxP connection in M = $\{m_{ij}\}$. In this model, there are two phases, recognition phase and learning phase. In the recognition phase, interactive association between input and output is executed until the convergence. In learning phase, the connection value is also updated during the interactive association (Eq 2-3). The activation value of each unit is an arbitrary continuous value in [-1, 1], and the teacher signal is given into output layer (to B). In learning phase, the dynamics is specified in the following.

$$\tau \dot{a}_i = - a_i + \Sigma_j S(b_j) m_{ij} \qquad (2\text{-}1)$$

$$\tau \dot{b}_j = - b_j + \Sigma_i S(a_i) m_{ij} + (T_j - S(b_j)) \qquad (2\text{-}2)$$

$$\tau_m \dot{m}_{ij} = - m_{ij} + S(a_i) S(b_j) \qquad (2\text{-}3)$$

$$(\text{for} \quad i = 1,....., N, \quad j = 1,....,P)$$

In this notation, $\dot{}$ is time differential, a_i and b_j is the inner state of input unit and output unit respectively, m_{ij} is connection value of between input and output, τ and τ_m are time constants of memory forgetting($\tau < \tau_m$), S(x) is monotonous increasing differential function like sigmoid function and decides the output value of each unit. On equation 2-2, the third term (T_j - $S(b_j)$) is a unique additional term in order to minimize recognition error.

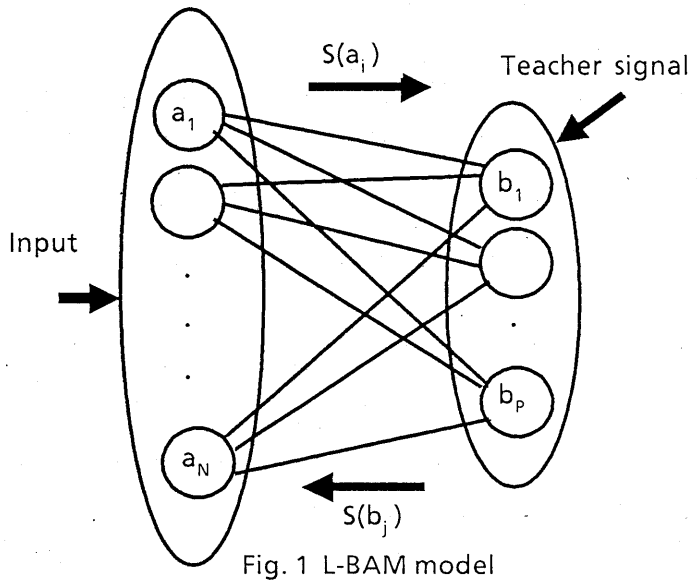

Fig. 1 L-BAM model

2.2 Lyapunov function and its stability

According to [1], I show that the following energy function E is converged on an local minimum of it without causing oscillation and chaos. In this approach, a combination of self-organization learning and least error learning with teacher is considered similar to energy learning method of Kawato[9] and self-organization with teacher of Sakaguchi[10].

$$E = (\text{self-organization energy}) + (\text{least error energy}) = E_S + E_T \tag{2-4}$$

$$E_S = \Sigma_i \int_0^{a_i} S'(x)\cdot x\, dx + \Sigma_j \int_0^{b_j} S'(x)\cdot x\, dx - \Sigma_i\Sigma_j S(a_i)S(b_j)m_{ij} + \Sigma_i\Sigma_j m_{ij}^2/2 \tag{2-5}$$

$$E_T = \Sigma_j (T_j - S(b_j))^2 / 2 \tag{2-6}$$

By the Eq (2-1) ~ (2-3), time differential of the energy function E is the following. In this notation, " · " shows time differential and " ' " shows space differential.

$$\begin{aligned}
\dot{E} &= \dot{E}_S + \dot{E}_T \\
&= \Sigma_i \dot{a}_i S'(a_i)a_i + \Sigma_j \dot{b}_j S'(b_j)b_j \\
&\quad - \Sigma_i\Sigma_j \{ \dot{a}_i S'(a_i)S(b_j)m_{ij} + \dot{b}_j S'(b_j)S(a_i)m_{ij} + S(a_i)S(b_j)\dot{m}_{ij} \} \\
&\quad + \Sigma_i\Sigma_j m_{ij}\dot{m}_{ij} - \Sigma_j (T_j - S(b_j))\dot{b}_j S'(b_j) \\
&= - \Sigma_i \dot{a}_i S'(a_i) \{ - a_i + \Sigma_j S(b_j)m_{ij} \} \\
&\quad - \Sigma_j \dot{b}_j S'(b_j) \{ - b_j + \Sigma_i S(a_i)m_{ij} + (T_j - S(b_j)) \} \\
&\quad - \Sigma_i\Sigma_j \dot{m}_{ij} \{ - m_{ij} + S(a_i)S(b_j) \} \\
&= - \tau \Sigma_i S'(a_i)\dot{a}_i^2 - \tau \Sigma_j S'(b_j)\dot{b}_j^2 - \tau_m \Sigma_i\Sigma_j \dot{m}_{ij}^2 \leqq 0
\end{aligned} \tag{2-7}$$

The lyapunov function is stable with $\dot{a}_i^2 = \dot{b}_j^2 = \dot{m}_{ij}^2 = 0$, since E is obviously bounded. The stable point depends on the initial values of input inner state and connection value. Each local minimum stable point of E shows respectively recognition state in the memory.

2.3 Interpretation of each term of the energy function

Each term of the energy function is interpreted by the following. Obviously, E_T in Eq (2-6) shows least error energy. According to Eq (2-1) ~ (2-3) and assuming S(x) as sigmoid function, the first and second terms of E_S in Eq (2-5) show forgetting Short Term Memory (STM) of each element, so as to make the inner state of each element to zero. The third term of them shows competitive-coorperative term between each input- output element and the connection, and the forth term of them shows forgetting Long Term Memory (LTM) in the connection not to emit the value.

Actually we must decide a rate of self-organization energy E_S to least error energy E_T. In this case, the Eq (2-2) and (2-4) are updated by the following.

$$E = E_S + \beta E_T \qquad (2\text{-}4')$$

$$\tau \dot{b}_j = - b_j + \Sigma_i S(a_i) m_{ij} + \beta(T_j - S(b_j)) \qquad (2\text{-}2')$$

β should be decreased as the learning process. At the early stage of learning, β should be large to have a strong effect of least error learning, but as the later stage near the convergence, β should be smaller and smaller to have the effect of minute adjustment by self-organization. This method is mathematically similar to Penalty Function Method.

3. Learning algorithm

After this section, the notation of variable a_i, b_j, m_{ij} is changed to I_i, O_j, w_{ij} respectively.

3.1 Learning algorithm

Step0: Setting of parameters

Set the following parameters to proper value.

learning rate: $\beta > 0$, variation of β: $0 < \beta' < 1$, recognition accuracy: $\varepsilon > 0$, time constant of forgetting: τ, τ_m

Step1: Setting of each input pattern

Set each input pattern $S(I^k)$ in order, and initialize output inner value to zero. Iterate Step2 - Step6 for $k = 1, ..., K$ respectively.

Step2: Main process

Each inner state and connection value is sychronously updated as following.

$$\Delta I_i = \{ - I_i + \Sigma_{j=1}^{P} S(O_j) w_{ij} \} \Delta t / \tau \qquad (3.1)$$

$$\Delta O_j = \{ - O_j + \Sigma_{i=1}^{N} S(I_i) w_{ij} + \beta(T_j - S(O_j)) \} \Delta t / \tau \qquad (3.2)$$

$$\Delta w_{ij} = \{ - w_{ij} + S(I_i) S(O_j) \} \Delta t / \tau_m \qquad (3.3)$$

$$I_i = I_i + \Delta I_i \qquad (3.4)$$

$$O_j = O_j + \Delta O_j \qquad (3.5)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad (3.6)$$

for all i, j

Step3: Decision of convergence with teacher signal

If $E_T = \Sigma_{j=1}^{P} (T_j - S(O_j))^2 / 2P \geqq \varepsilon$, then go to Step2.

Step4: Stop criterion

If β is very small (< 1.0) or there is no error recognition in all K-th patterns as the recognition phase[note], stop the algorithm.

Otherwise, set $\beta = \beta \times \beta'$, and go to Step1.

note): In the recognition phase, interactive association between input and output is executed sychronously without the third term of Eq (3-2) , Eq (3-3) and Eq (3-6) as least error learning. The stable convergency of recognition process have already proved by B. Kosko [1].

3.2 Setting of the initial connection value

The initial connection value is set by each clustering center in order that output may represent a distance (similarity) between an input pattern and each clustering center as prototype pattern like SPAN [7]. Each clustering center is normalized with constant value $\sqrt{N}$ to equalize inner product distance to Euclid distance. Since the clustering center expression <Case1> is better than the conventional correlation matrix expression <Case2> in regard to the influence of crosstalk, Case1 is adopted.

<Case1 clustering center expression >

The distance between an input pattern S(I) and each clustering center C_j (j=1, ..., P) is shown by the following (T means transpose).

$$d_j^2 = (S(I) - C_j)^T (S(I) - C_j) = 2N (1 - C_j{}^T S(I) / N)$$

$$O_j \equiv (1 - d_j 2 / 2N) = C_j{}^T S(I) / N = \Sigma_{i=1}^{N} S(I_i) C_{j\,i} / N = \Sigma_{i=1}^{N} S(I_i) w_{ij}$$

$$\therefore\ w_{ij} = C_{j\,i} / N \qquad (3.7)$$

<Case2 correlative matrix expression>

As in case1, output is considered to represent the distance of an input and clustering centers, but connection value is set by correlation matrix, different from case1.

$$W \equiv \Sigma_{k=1}^{K} S(O^k) (S(I^k))^T / N$$

$$S(O^k_j) \equiv C_j{}^T S(I^k) / N$$

are assumed. When key pattern $S(I^l)$ is given to input layer, then output is

$$\begin{aligned} O_j &= \Sigma_{i=1}^{N} S(I^l_i) w_{ij} \\ &= \Sigma_{i=1}^{N} S(I^l_i) \Sigma_{k=1}^{K} S(O^k_j) (S(I^k_i)) / N \\ &= \Sigma_{k=1}^{K} C_j{}^T S(I^k) \Sigma_{i=1}^{N} S(I^l_i) S(I^k_i) / N^2 \\ &= C_j{}^T S(I^l) / N + \Sigma_{k \neq l} C_j{}^T S(I^k) \Sigma_{i=1}^{N} S(I^l_i) S(I^k_i) / N^2 \end{aligned} \qquad (3.8)$$

The first term of Eq (3.8) shows the distance between an input and clustering centers as case1, and the second term shows the crosstalk noise from other patterns. In case1, the distance between input and only own clustering center influences the inner state of elements at feedforward process, and the distance between input and other clustering centers influences the inner state of elements at backward process.

4. Simulation of character recognition

Through a simulation of character recognition identify the parameter properties of L-BAM, L-BAM model is compared with BP model in recognition accuracy and learning iteration time.

4.1 Objects for recognition

In this simulation, hand-written Japanease HIRAGANA characters made by digitizer are used as objects. Five characters " あ", " い", " う", " え", " お", which have five variations respectively, by 16x16 dotted pattern (± 0.95) is used. The dimension of input layer is 256 and output layer is 5.

4.2 Method of simulation

L-BAM and BP are compared in the learning iteration time and the data complementary ability by changing the parameters for the hand-written character in both cases of using only one character and all the patterns for learning. The simulation environment is constructed on Sun3 by using neuro-simulator SunNet, which is public domain neuro-simulator software for Sun3 developed by Dr. Miyata at UCSD (he is in AT&T Bellcore).

4.3 The simulation result of L-BAM (as shown in Fig.2)

Before learning, the similarity once grows, but converges on a constant value due to hysteresis; however after learning, the pattern can be recognized[12][13]. As the results, the following points are identified.

i) The variation β' shows a kindness of guidance. Without the case of unkind guidance (small β') at Fig.2(a), learning iteration time does not almost depend on β'.

ii) A large number of learning is needed in the case of large τ_m (memory dependence is strong) and small β' (unkind guidance) at Fig.2(b).

iii) As the forgetting of LTM is larger (τ_m is smaller), the learning iteration time becomes fewer.

4.4 The simulation results of BP (as shown in Fig.3)

The learning iteration time is investigated changing the learning step width:η and momentum:α. Each parameter is set as large as with the range of no vibration [14]. A convergence is defined as recognition error smaller than 0.01, when the learning process is terminated. The initial connection value is same as L-BAM at input-hidden, but connection of hidden-output is set by correlation matrix between the recognition output pattern and the similarity as distance for each cluster.

iv) The learning iteration time is needed 10 times more than L-BAM, and depends more strongly on the parameters than L-BAM.

v) A few patterns are always not able to recognize correctly in the case of using only one variation character in learning.

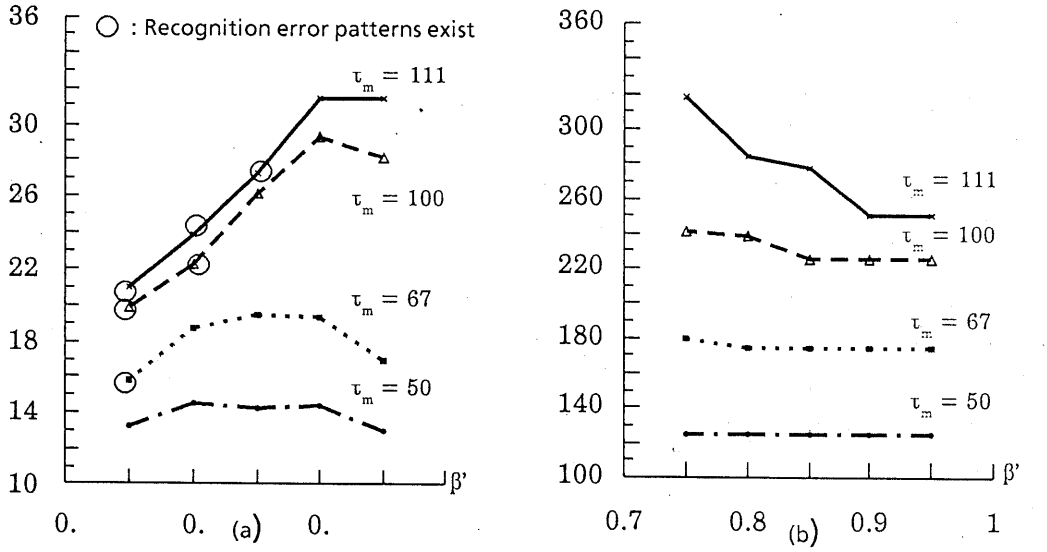

Fig. 2 Learning iteration times vs. variation β' (L-BAM)
β = 128, ε = 0.001, Δt = 0.1, τ = 10, $S(x) = 2/(1 + e^{-10x}) - 1$.

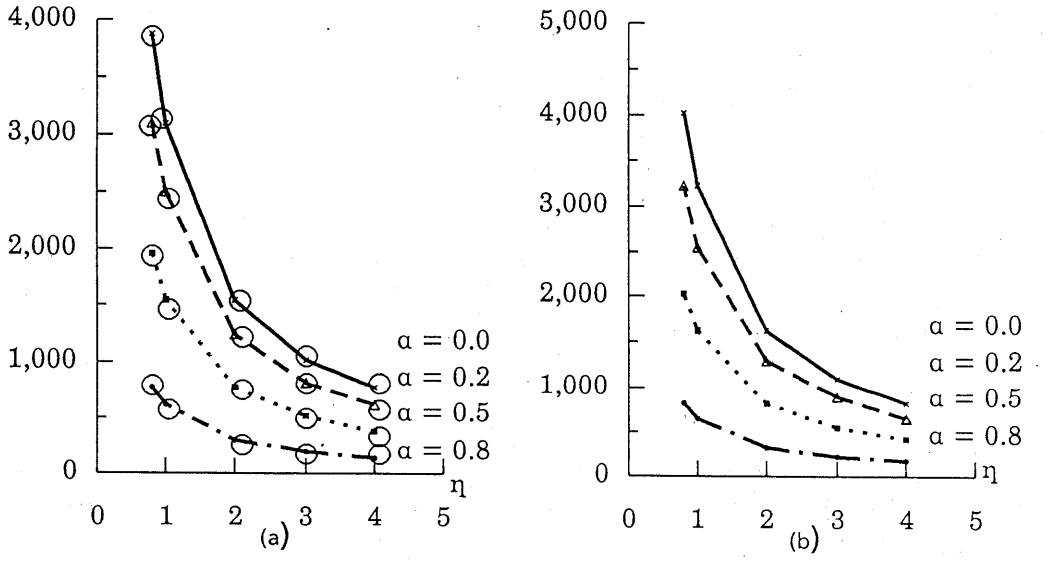

Fig. 3 Learning iteration times vs. step width η (BP)
In both Fig.2 and Fig.3, (a) is the case of only one variation character for learning, and (b) is the case of using all pattern for learning.

5. Consideration

The time constant τ_m controls the relative the synaptic plasticity in changing the value of LTM. There are two bipolar cases, one has a tendency for forgetting when τ_m is small and the other has a tendency for stabilizing when τ_m is large. Proper intermediate values of τ_m are important to proper operation. In a micro interpretation of physiology, the least-square error learning term of L-BAM can be considered as a hetero-synaptic learning term. In a macro interpretation, the input-ouput interaction might be considered as one of models of the corpus callosum [15].

Several Neural Network models are described in [8][16] including BP and VQ. BP and VQ use two-dimensional clustering while L-BAM uses three-dimensional clustering. L-BAM has also nonlinear clustering boundaries. So, if the initial distance is small, the state of output approaches the correct output in the basin of the attractor. We are investigating the ability of generalization of this technique based on the basin of the attractor. Early results on small amounts of test data are encouraging. We plan to test on a wide variety of data in the future.

Acknowledgments.

I would like to thank Dr. Asou in ETL and Dr. Miyata in AT&T for many answers to my questions on the use of SunNet.

References.

1) B.Kosko: Adaptive bidirectional associative memories, Applied Optics, Vol.26, No.23 pp.4947-4960 (1987).

2) B.Kosko: Feed Back Stability and Unsupervised Learning, ICNN '88 Proceedings (1988).

3) K.Nakano: Associatron, Shoukidou printed in Japan (1979).

4) T.Kohonen: Associative Memory A System - Theoretical Approach, Springer - Verlag (1977).

5) J.J.Hopfield: Neural networks and physical systems with emergent collective computational abilities, Proc. Nat Acad Sci. USA Vol.79, pp.2554-2558 (1982).

6) K.Funahashi: On the Approximate Realization of Continuous Mappings by Neural Networks, Neural Networks, Vol 2, No.3, pp. 183- 192 (1989).

7) H.Kawahara and T.Irino: A Procedure for Designing 3-Layer Neural Networks Which Approximate Arbitrary Continuous Mapping: Application to Pattern Processing, IEICE PRU 88-54, pp. 47-54, in Japan (1988).

8) S.Katagiri: Systematic Explanation of Learning Vector Quantization and Multi-layer Perceptron - Proposition of Distance Network -, IEICE MBE 88-72, pp.75-82, in Japan (1988).

9) M.Kawato et al. : MULTI-LAYER NEURAL NETWORK MODEL WHICH LEARNS AND GENERATES HUMAN MUTI-JOINT ARM TRAJECTORY , IEICE MBE 87-133, pp.223-240 in Japan (1987).

10) Y.Sakaguchi: Topographic Organization of Nerve Field with Teacher Signal, IEICE-DII No.5, pp.782-791, in Japan (1989).

11) K.Nakano et al. : Model of Neural Visual System with Self-Organizing Cells, Biol Cybern 60, pp.195-202 (1989).

12) S.Amari: Statistical Neurodynamics of Various Versions of Correlation Associative Memory, ICNN '88 Proceedings (1988).

13) S.Amari and K.Maginu: Statistical Neurodynamics of Associative Memory, Neural Networks, Vol 1, No.1, pp.63-73 (1988).

14) K.Yamada et al. : Character Recognition using Neural Network, IEICE PRU 88-58, pp. 79-86, in Japan (1988).

15) N.D.Cook: THE BRAIN CODE Mechanisms of information Transfer and the Role of the Corpus Callosum (1986)

16) R.P.Lippmann: An Introduction to Computing with Neural Nets, IEEE ASSP Magazine (1987).