Feedback Loop を持つ Boltzmann Machine によるパターン遷移の学習

5F-6

野田 五十樹, 長尾 真
京都大学工学部

1 はじめに

神経回路網には「類似したパターンに対しては類似した処理が行なわれる」という特徴がある。この特性を有効に利用するためには、行う処理に対して適切なパターン表現を選ぶ必要がある。本稿ではこのパターンの類似性と処理の類似性の対応関係を定量的に表わし、行なう処理にあった内部パターンを生成する方法とその実験結果を示す。

2 モデル

本稿では、「ある規則にしたがってパターンを遷移させること」を処理と考える。そして、このパターンの遷移を神経回路網の学習対象とする。そこで、神経回路網のモデルとして、図1に示す feedback loop を持つ Boltzmann Machine [2] [3] を採用した。

H^-層は feedback によって H層の1時刻前の状態を保持している層である。V層は外界とのインターフェースであり、ここに提示される外部パターンは V-H間の connection $W^V(=[w^V_{ij}])$ により H上の内部パターンに変換する。また、H^--H間の connection $W^T(=[w^T_{ij}])$ で内部パターンの遷移規則を表現する。$W^S(=[w^S_{ij}])$ は H層の層内結合である。モデルの基本的な動作,及び connection の weight の学習は Boltzmann Machine の動作原理, 学習則 [2] に基づいて行う。今後、V,H,H^-上のパターンをベクトル表記 $\xi(=[\xi^i]),\mu(=[\mu^i]),\lambda(=[\lambda^i])$ で表わすものとする。

Boltzmann Machine の動作は確率的であるので、本モデルの H上のパターンの遷移は確率的遷移である。よって、ここでは H上のパターンの遷移確率を処理規則と考える。3節ではこの遷移確率と H^-上のパターンとの関係について述べ、学習が効率的に行える内部パターンの条件を調べる。さらに、4節では Vから Hへのパターン変換について述べる。

3 遷移確率とパターンの関係

本モデルでは、任意の内部表現に対して任意の遷移確率を実現できるわけではなく、実現できる遷移確率はその内部表現にかなり依存する。そこで、この依存関係を定量的に表わすことを試みた。解析を容易にするために V層の影響を無視して議論を進める。

Boltzmann Machine の動作原理により H^-上の パターン が λ である時、H上に パターン μ が生じる確率 (以後、"λ に於ける遷移確率"と略す) は

$$p(\mu|\lambda) = e^{w^T_{ij}\lambda^i\mu^j + w^S_{ij}\mu^i\mu^j + h_i\mu^i}/Z_\lambda \qquad (1)$$

で表すことができる。ただし、Z_λ は正規化定数である。いま、ある パターン λ が $\{\lambda_i|i=1\ldots M\}$ によって

$$\lambda = \sum_i a_i\lambda_i + d\lambda, \qquad \sum_i a_i = 1 \qquad (2)$$

で誤差 $d\lambda$ で近似できるとする。このとき、各々のパターンに於ける遷移確率 $p(\mu|\lambda),p(\mu|\lambda_i)$ の間には、$d\lambda$ が無視できるならば次の関係が成立する。

$$p(\mu|\lambda) = K_\lambda \prod_i (p(\mu|\lambda_i))^{a_i} \quad = p'(\mu) \qquad (3)$$

ただし、K_λ は正規化定数である。すなわち、内部パターン同士に 式(2) の関係があるときには、式(3) を満たさないような遷移確率は実現できない。しかし、内部パターンを線形独立であるように選ぶと、$d\lambda$ は 0 ではなく、このような場合、任意の遷移確率が実現できることが知られている [1]。ところが、実際に学習させてみると、線形独立なパターンでも学習が困難なものが存在する。そこで、これを $d\lambda$ と 式(3) の関係で表わすことを試みる。いま、次のような仮定が成立するとする。

- w^T_{ij} の値の分布は平均 $\bar{w}^T$, 分散 $\sigma^2_{w^T}$ の正規分布 $N(w^T;\bar{w}^T,\sigma^2_{w^T})$ をなす。
- $d\lambda^i\mu^j$ と w^T_{ij} は無相関であり、$w^T_{ij}d\lambda^i\mu^j$ の値の分布は 、$n = \|d\lambda\|^2\ (=\sum_i d\lambda^{i2}), m = \|\mu\|^2$ とすれば、平均 $nm\bar{w}^T$, 分散 $nm\sigma^2_{w^T}$ の正規分布 $N(w^T_{ij}d\lambda^i\mu^j; nm\bar{w}^T, nm\sigma^2_{w^T})$ をなす。
- 推定遷移確率 $p'(\mu)$ を $p = e^{x_\mu}$ で表わすとすると、x_μ と $w^T_{ij}d\lambda^i\mu^j$ は無相関である。

この時、実際の遷移確率 $p(\mu|\lambda)$ と式(3) により推定される遷移確率 $p'(\mu)$ との差 Kullback Divergence $G(p;p')$ の期待値は次のように表わされる。

$$\langle G(p;p')\rangle = nm\sigma^2_{w^T}/2 \qquad (4)$$

すなわち、$d\lambda$ が無視できないときには式(3) の等式に式(4) で示されるような誤差が生じ、その誤差内であるような遷移確率は実現が容易であると予想できる。

この結果を用いて学習の困難さを表わすことを試みる。いま、パターン $\lambda_i(i=1\cdots M)$ に於ける遷移確率 $p(\mu|\lambda_i)$ が connection W^T,W^S によって、実現されているとする。ここへ未知パターン λ に於ける遷移確率 $p(\mu|\lambda)$ を学習させる場合、λ を既知パターンλ_iによって式(2) のように近似したときに 式(3) から得られる推定遷移確率 p'と、学習によって与えられる遷移確率 p との差 $G(p;p')$ は、学習が完了した時点で式(4) を満たしていると考えることができる。ここで、n,m が一定であるので、$G(p;p')$ が大きいときには $\sigma^2_{w^T}$ が大きくなる。$\sigma^2_{w^T}$ の大きさは、この場合、学習に於ける connection の weight の修正量を表わすので、weight の修正量が大きいほど学習が難しいとすると、$G(p;p')/nm$ は学習の困難さを表わす指標とみることができる。

この指標と学習の精度 (1000 回学習後の遷移確率 q と与えられる遷移確率 p の差 $G(q;p)$ の平均) との関係を、具体的なパターンと遷移確率について計算機による実験によって調べた。その結果を 図2 に示す。これによれば、この指標は学習の難しさをある程度反映していることがわかる。

Learning of Pattern Transition by Boltzmann Machine with Feedback Loop
Itsuki NODA and Makoto NAGAO
Kyoto University

4 パターン変換

3節で述べたように、H上のパターンは実現できる遷移規則を制限するので、W^V によるパターン変換はシステム全体の能力を決定する上で重要な役割を持つ。この W^V による Vから Hへのパターン変換は確率的に行なわれるため、情報損失が生じる。この情報損失が大きければ H^--H間の正しい遷移確率を学習することはできない。そこで、まず、H^-の影響を無視できるものとしてこのパターン変換による情報損失と W^V,W^Sの統計量の関係を調る。いま、外部から Vにパターン$\boldsymbol{\lambda}$ が与えられる確率を $q(\boldsymbol{\xi})$ とし、$x=(w^V_{ij}\xi^i\mu^j)/T, y=(w^S_{ij}\mu^i\mu^j+h_i\mu^i)/T$ とすると、情報損失量は

$$H(V|H)=\sum_{\boldsymbol{\xi},\boldsymbol{\mu}} q(\boldsymbol{\xi})\frac{\exp(x+y)}{Z_{\boldsymbol{\xi}}}\log\left(\frac{\sum_{\boldsymbol{\xi}'} q(\boldsymbol{\xi}')\frac{\exp(x+y)}{Z_{\boldsymbol{\xi}'}}}{q(\boldsymbol{\xi})\frac{\exp(x+y)}{Z_{\boldsymbol{\xi}}}}\right) \quad (5)$$

であらわせる。このとき、x,y の値の分布が、平均0, 分散σ_x^2,σ_y^2の正規分布 $\boldsymbol{N}(x;0,\sigma_x^2)$,$\boldsymbol{N}(y;0,\sigma_y^2)$ をなすと仮定し、計算機により具体的な数値で情報損失を求めると図3のようになる。これによれば、情報損失の面からみれば σ_y^2/σ_x^2 はできるだけ小さい方がよく、そのためには w^S_{ij},w^T_{ij} の分散は小さい方がよい。同様に、Hが Vに対し十分大きくすると σ_y^2/σ_x^2 が大きくなり、情報損失のために外部より与える処理が学習できなくなることがわかる。

次にパターン変換について考える。上の議論より、情報損失の点からみると、学習を効率的に行なうためには、W^V の学習を W^S,W^T の学習の前に行なっておく、あるいは W^S にランダムな初期値を与えておく方法が良いと考えられる。しかし、これらの方法では H上のパターンは V上のパターンの類似性のみをを反映したものとなり、V上のパターン同士に 式(2) のような関係も H上のパターンに反映されやすくなる。よって、実現できる遷移規則は 式(3) の関係を満たすものに限られてしまう。V上のパターンの類似性に依存しないような処理を実現するためには、H上のパターンが H^-上のパターンにも依存した形で生成されるようにしなければならない。このためには **W^T の学習は W^V と同時に行なえばよい。**

以上の議論を実証するために、以下のような実験を行なった。まず V層に提示するパターンとして式(2) の意味で線形従属なパターンを選び、それに対し、式(3) を満たさないような遷移確率を与える。このような学習セットを (a) W^V の学習をさきに行なうもの (b) W^T を W^V と同時に学習するものの2種類のモデルについて学習を行なわせた。結果を 図4 に示す。この結果から、最終的には (b) の方が学習精度が良くなることがわかった。これにより、W^T を W^S と同時に学習させる方が効率がよいことがわかった。しかし、上で述べているように Hが Vに比べ大きい場合には情報損失が減少せず、学習は失敗してしまう。これを避けるためには、なんらかの方法で σ_y^2/σ_x^2 を 1 以下に保つようにすることが良いと思われる。

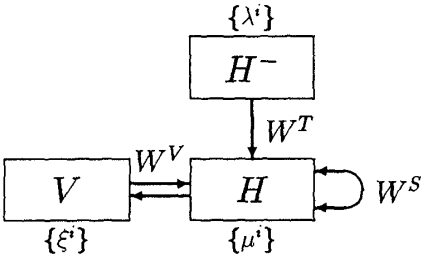

図 1: feedback loop を持つ Boltzmann Machine

5 おわりに

本稿では feedback loop を持つ Boltzmann Machine についてその内部パターンとパターン処理の能力, 学習の困難さの関係を定量的に表わし、それをもとにパターンの遷移確率の学習の困難さの指標を提案し、実験によりそれを評価した。また、内部パターンへのパターン変換について、情報損失、処理の類似性の観点から考察を行ない、それを元に処理に適した内部パターンを生成する方法を提案した。今後はこれを応用し、より複雑なパターン処理を効果的に学習する方法を構成して行く予定である。

参考文献

[1] 倉田, 甘利. 確率的に動作する自己組織神経回路網について. 技術研究報告 MBE85-104, 電子情報通信学会, 1986.

[2] Ackley D.H. et.al. A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[3] R.W. et.al. Prager. Boltzmann machine for speech recognition. *Computer Speech and Language*, 1:3–27, 1986.

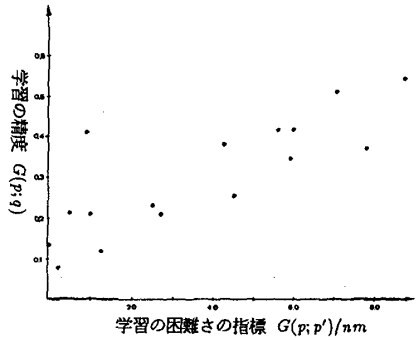

図 2: 学習の困難さの指標と学習

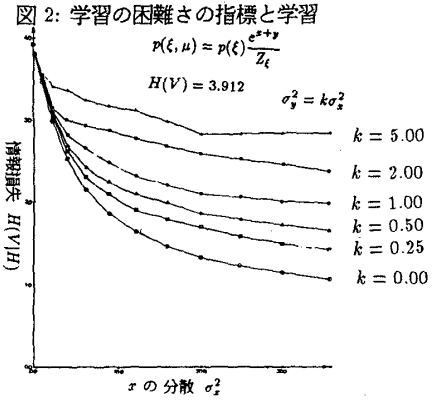

図 3: weight の分散と情報損失

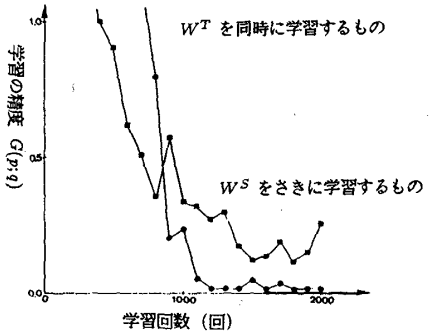

図 4: weight の学習のタイミングと学習の精度