非単調（閾値関数をもつ）ネットワークの汎化能力

井上純一[1] 、西森秀稔（東工大理)、樺島祥介（東工大総合理工）

ニューロンの入出力特性が非単調であるようなネットワーク [1,2,3] は記憶容量の増大、疑記憶の識別等、工学的に好ましい性質を持つため、その詳しい特性が数理的に解析されている。これらの多くは、相互結合型ネットワークがp個のランダムパターン$\boldsymbol{\xi}^\mu$をヘップ則 $J_{ij}^{\mu+1}\to J_{ij}^{\mu}+\eta\xi_i^\mu\xi_j^\mu/N$ 等により学習した場合、p個のパターン各々がネットワークの状態$\boldsymbol{s}$の安定な固定点となるための臨界パターン数や、固定点に至るダイナミックス、引き込み領域等に関するもので、問題の視点は$\boldsymbol{J}$（学習）のダイナミックスよりも$\boldsymbol{s}$（ネットワークの状態）のダイナミックス及びその平衡状態の解明に向けられてきた。しかし、非単調ネットワークがどの様に外界の信号に適応し、結合を変えていくのかという動的問題、特に学習に用いる例題がランダムパターンでなく、非単調ネットワークの入出力の場合の教師有り学習を考え、その時の汎化能力を評価することは極めて重要である。本研究では、内部電位 $u=\sqrt{N}(\boldsymbol{J}\cdot\boldsymbol{x})/|\boldsymbol{J}|$ （$\boldsymbol{x}$は入力ベクトル）に対する入出力が $S_a(u)=\mathrm{sign}[u(a-u)(a+u)]$ で与えられる非単調ネットワークが、同じ構造を持つ教師ネットワーク $T_a(v)=\mathrm{sign}[v(a-v)(a+v)]$ （$v=\sqrt{N}(\boldsymbol{B}\cdot\boldsymbol{x})/|\boldsymbol{B}|$）から逐次型学習する場合を調べる。逐次型学習では、各例題は参照後に捨てられてしまうので、学習のダイナミックスは次のようになる

$$\boldsymbol{J}^{\mu+1}=\boldsymbol{J}^{\mu}+f(T_a(v),u)\boldsymbol{x}$$

これは教師、生徒機械の重なり $R^\mu=(\boldsymbol{B}\cdot\boldsymbol{J}^\mu)/|\boldsymbol{B}||\boldsymbol{J}^\mu|$, 生徒機械の結合のノルム $l^\mu=|\boldsymbol{J}^\mu|/\sqrt{N}$ を用いると、$N,p\to\infty$、$\alpha\equiv p/N\sim\mathcal{O}(1)$ で次の方程式になる

$$\frac{dl}{d\alpha}=\ll f^2(T_a(v),u)+2f(T_a(v),u)ul\gg/2l$$

$$\frac{dR}{d\alpha}=\ll-\frac{R}{2}f^2(T_a(v),u)-(Ru-v)f(T_a(v),u)l\gg/l^2$$

ここで、$\ll\cdots\gg$は、例題に関する平均。代表的な学習の重み関数として、(I) ヘップ：$f=-T_a(v)$ (II) パーセプトロン：$f=-S_a(u)\Theta(-T_a(v)S_a(u))$ (III) アダトロン $f=-u\Theta(-T_a(u)S_a(v))$ があり、これらの学習則の優劣は汎化誤差$\epsilon_g=\ll\Theta(-T_a(v)S_a(u))\gg$ によって測られる。$a\to\infty$ の場合は単調パーセプトロン同志の学習となり、この場合には例題数無限大の漸近領域で (I) $\epsilon_g\sim\alpha^{-1/2}$ (II) $\epsilon_g\sim\alpha^{-1/3}$ (III) $\epsilon_g\sim\alpha^{-1}$ といずれもゼロに収束する。a が有限の場合でも、生徒、教師両ネットワークの構造が同じであることから、例題数が無限大の極限では$\boldsymbol{J}\to\boldsymbol{B}$ となり、誤差ゼロとなることが予想される。しかし、実際には (I)(II) の学習則を用いた場合、$a<a_c=\sqrt{2\log 2}$を満たすようなパラメータ領域では汎化誤差は驚くべきことに単調に 1 に向かって増加してしまうことが解った。

この「アンチラーニング」は次のようにして確認できる。

例としてヘップ則 (I) により、1 つの例題を学習する場合を考えると、$a=0<a_c=\sqrt{2\log 2}$のとき

$$\boldsymbol{J}=T_{a=0}(v)\boldsymbol{x}$$

このとき、生徒機械の出力は

$$\begin{aligned}S_{a=0}(u) &= -\mathrm{sign}(u)\\ &= -\mathrm{sign}(\boldsymbol{J}\cdot\boldsymbol{x})\\ &= -T_{a=0}(v)\end{aligned}$$

[1]E-mail: jinoue@stat.phys.titech.ac.jp www: http://www.stat.phys.titech.ac.jp/hp/jinoue/

となり、教師$\boldsymbol{B}$とは逆方向に学習してしまうことが解る。このアンチラーニングを解消するために、我々は幾つかの学習則を提案した。ここでは、その内の1つを紹介したい。

学習ができるだけ速く完了させるためには、$\boldsymbol{J}$が$\boldsymbol{B}$に素早く一致させればよいが、このためにRの変化率$dR/d\alpha$を各ステップで最大化させるように学習の重み関数fを選ぶことを考える。このfは生徒機械には知り得ないパラメータvを含んでいるが、これをベイズ統計を用いて最適化する。こうして、利用できる情報を全て用いた場合、Rの時間発展の方程式は次のようになる

$$\frac{dR}{d\alpha} = \frac{1-R^2}{4\pi R}\int_{-\infty}^{\infty} Du\, \Omega_a(R,u)$$

この場合、汎化誤差は漸近的にaに依らず

$$\epsilon_g = \frac{2\sqrt{2\pi}}{\int_{-\infty}^{\infty} dt \exp(-t^2)/H(t)}\frac{1}{\alpha} \sim \frac{0.883}{\alpha}$$

となるが、これは非逐次型学習でのバウンド

$$\lim_{\alpha\to\infty} \alpha\epsilon_g(\alpha) \leq 0.44$$

の2倍である。また、関数$\Omega_a(R,u)$は生徒の有効な質問についての情報を与える。$(1-R^2)/4\pi R \geq 0$であるから、ある学習過程にRにおいて、関数Ω_aの最大値を与えるuを$u_{\max}$とすると、$u_{\max} = \sqrt{N}(\boldsymbol{J}\cdot\boldsymbol{x}_{\max})/|\boldsymbol{J}|$となるような質問$\boldsymbol{x}_{\max}$が有効的であり、この時の汎化能力は

$$\frac{dR}{d\alpha} = \frac{1-R^2}{4\pi R}\int_{-\infty}^{\infty} Du\, \delta(u-u_{\max})\, \Omega_a(R,u)$$

により与えられる。より詳しい解析や、紙面の都合上ここでは割愛せざるを得なかった他の学習則に関しては文献[4]を参照されたい。

また最近では、同じ例題が繰返し選ばれて与えられてしまう場合の、所謂「不完全訓練データ」からの学習についても議論され始めており、今後の課題として残されている[5]。

謝辞：井上はこの研究を進めるにあたり理化学研究所ジュニア・リサーチ・アソシエイト制度により一部援助を受けました。

---

[1] M. Morita, S. Yoshizawa and K. Nakano, *Trans.Inst. Electron. Inf. Commun.* **J73-D-II**, 242 (1993).
[2] G. Boffeta, R. Monasson and R. Zecchina, *J.Phys.A: Math.Gen.* **26**, L507 (1993).
[3] H. Nishimori and I. Opris, *Neural Networks* **6**, 1061(1993).
[4] J. Inoue, H. Nishimori and Y. Kabashima, Submitted to *Phys.Rev.E* (cond-mat/97051920).
[5] C. W. H. Mace and A. C. C. Coolen, *King's College London preprint*, (1997).