教師がコミティマシンの場合のアンサンブル学習

Analysis of Ensemble Learning for Committee Machine Teacher

三好 誠司 (P)†, 原 一之‡, 岡田 真人*

Seiji Miyoshi(P), Kazuyuki Hara and Masato Okada

† 神戸高専 電子工学科, ‡ 都立工業高専 電子情報工学科, * 東大院, 理研 脳総研, 科技機 さきがけ

miyoshi@kobe-kosen.ac.jp

Abstract— Ensemble learning, in which a teacher and students are a committee machine and simple perceptrons respectively, is discussed based on online learning theory. Perceptron learning keeps a variety of students and the effect of ensemble doesn't disappear. Dynamics of generalization error by AdaTron learning shows the minimum point.

Keywords— **ensemble learning, online learning, committee machine, generalization error**

1 まえがき

精度の低いルールや学習機械（以後は生徒と呼ぶ）を複数組み合わせることにより精度の高い予測や分類を行おうとすることは一般にアンサンブル学習と呼ばれ，近年注目されている [1, 2, 3]．アンサンブル学習の汎化能力を統計力学的手法によって理論的に解析する研究もさかんに行われている [4, 5, 6, 7]．

著者らは教師が単純パーセプトロンで生徒が K 個の単純パーセプトロンであるようなアンサンブル学習を，オンライン学習の枠組みで議論した [6, 7]．その結果，単純パーセプトロンの学習則としてよく知られているヘブ学習，パーセプトロン学習，アダトロン学習 [8, 9] の三学習則が「生徒の多様性維持」というアンサンブル学習との相性という点でそれぞれ異なった性質を有しており，アダトロン学習が，アンサンブル学習との相性という点で最も優れているという興味深い事実が明らかになった．一方，Inoue と Nishimori は教師が一個の非単調パーセプトロンであり生徒が一個の単純パーセプトロンである場合について解析した [9]．Inoue と Nishimori が扱ったモデルは，教師が生徒のモデル空間内にない場合ということができる．

アンサンブル学習の大きな特徴として，多数決などで生徒を組み合わせることにより，単一の生徒では表現できない入出力関係を実現できることがあげられる [3]．その意味で，教師が生徒一個のモデル空間内にないような場合のアンサンブル学習の解析は非常に興味深い．そこで本論文では，教師がコミティマシンであり，生徒が単純パーセプトロンの集団であるようなアンサンブル学習についてオンライン学習の枠組みで議論する．その結果，パーセプトロン学習では生徒の多様性が消滅せず，そのためにアンサンブルの効果が残ること，アダトロン学習では汎化誤差がいったん最小になってその後少し増大することなど，興味深い事実が明らかになる．

2 モデル

生徒は符号関数を出力関数とするパーセプトロンである．K 個の生徒からなるアンサンブルを考え，各生徒の結合荷重 $\boldsymbol{J}_1, \boldsymbol{J}_2, \cdots, \boldsymbol{J}_K$ ，入力 $\boldsymbol{x}$ は N 次元ベクトルである．$\boldsymbol{J}_k$ の初期値 $\boldsymbol{J}_k^0$ の各要素 J_{ki}^0 は平均 0，分散 1 のガウス分布にしたがい独立に生成されるものとする．また，$\boldsymbol{x}$ の各要素 x_i は平均 0，分散 $1/N$ のガウス分布に従う独立な確率変数であるとする．各生徒の出力は $\mathrm{sgn}(u_1 l_1), \mathrm{sgn}(u_2 l_2), \cdots, \mathrm{sgn}(u_K l_K)$ である．ここで，$u_k l_k = \boldsymbol{J}_k^T \boldsymbol{x}$ である．l_k については後で述べる．u_k を各生徒の規格化内部状態と呼ぶことにする．

教師機械は中間層ユニット数 M のコミティマシンであるとする．各中間層ユニットは符号関数を出力関数とするパーセプトロンである．入力層から各中間層への結合荷重 $\boldsymbol{B}_m$ は N 次元ベクトルであり，その各要素 B_{mi} は平均 0，分散 1 のガウス分布にしたがい独立に生成され，不変であるとする．$v_m = \boldsymbol{B}_m^T \boldsymbol{x}$ を教師中間層の内部状態と呼ぶことにする．教師出力ユニットは中間層ユニット出力の単純多数決をとる．すなわち教師の出力は $d = \mathrm{sgn}(\sum_{m=1}^{M} \mathrm{sgn}(v_m))$ である．本論文では，$N \to \infty$ の熱力学的極限を考えることにする．このとき $|\boldsymbol{x}| = 1$, $|\boldsymbol{B}_m| = |\boldsymbol{J}_k^0| = \sqrt{N}$ となる．生徒の大きさ $|\boldsymbol{J}_k|$ は一般には時間の経過とともに変化するが，$\sqrt{N}$ に対する比を l_k とし，これを生徒 $\boldsymbol{J}_k$ の長さと呼ぶことにする．

教師機械と個々の生徒には共通の入力 $\boldsymbol{x}$ が同じ順序で与えられる．個々の生徒は入力 $\boldsymbol{x}$ に対する教師の出力と自分の出力を比べ，教師と同じ出力を出す確率が上がるように，必要に応じて自分の結合荷重を修正していく．この手続きを学習と呼ぶ．修正の方法は学習則と呼ばれ，ヘブ学習，パーセプトロン学習，アダトロン学習がよく知られている [8, 9] ．自分自身に関する情報以外に生徒が修正のために使える情報は，入力 $\boldsymbol{x}$ とそれに対する教師の出力 d だけであるから，学習は一般に $\boldsymbol{J}_k^{n+1} = \boldsymbol{J}_k^n + f\left(d^n, u_k^n\right) \boldsymbol{x}^n$ と書ける．ここで n は時刻ステップを表す．

3 理 論

3.1 汎化誤差

統計的学習理論の目的のひとつは汎化誤差 ϵ_g を理論的に求めることである．本論文では K 個の生徒は多数決で集団としての出力を決定する場合を考える．このとき，誤差 ϵ として，$\epsilon = \Theta\left(-d \sum_{k=1}^{K} \mathrm{sgn}\left(\boldsymbol{J}_k^T \boldsymbol{x}\right)\right)$ を用いることにする．汎化誤差 ϵ_g は ϵ を入力 $\boldsymbol{x}$ の確率分布 $p(\boldsymbol{x})$ で平均したものと定義する．ϵ は教師中間層の内部状態 v_m と生徒の規格化内部状態 u_k を用いて $\epsilon = \epsilon(\{v_m\}, \{u_k\})$ と書くことができるので，ϵ_g も v_m，u_k の確率分布 $p(\{v_m\}, \{u_k\})$ を用いて，

$$\epsilon_g = \int \prod_{m=1}^{M} dv_m \prod_{k=1}^{K} du_k p(\{v_m\}, \{u_k\}) \epsilon(\{v_m\}, \{u_k\}) \quad (1)$$

と書ける．v_m と u_k は入力 $\boldsymbol{x}$ とそれとは無関係な変数 $\boldsymbol{B}_m$, $\boldsymbol{J}_k$ で書けるので $p(\{v_m\}, \{u_k\})$ は平均 0 の多重ガウス分布である．$p(\{v_m\}, \{u_k\})$ の共分散行列 $\boldsymbol{\Sigma}$ は，教師中間層 $\boldsymbol{B}_m$ と生徒 $\boldsymbol{J}_k$ の方向余弦 R_{mk} と生徒 $\boldsymbol{J}_k$ と生徒 $\boldsymbol{J}_{k'}$ の方向余弦 $q_{kk'}$ で書くことができる．すなわち，汎化誤差 ϵ_g は R_{mk} と $q_{kk'}$ を用いて以下のように書ける．ここで $\boldsymbol{I}$ は $M \times M$ の単位行列である．

$$p(\{v_m\}, \{u_k\}) = \frac{1}{(2\pi)^{\frac{M+K}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \times \exp\left(-\frac{(\{v_m\}, \{u_k\}) \boldsymbol{\Sigma}^{-1} (\{v_m\}, \{u_k\})^T}{2}\right) \quad (2)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\Sigma}_B \\ \boldsymbol{\Sigma}_B^T & \boldsymbol{\Sigma}_D \end{pmatrix}, \boldsymbol{\Sigma}_B = \begin{pmatrix} R_{1,1} & \cdots & R_{1,K} \\ \vdots & \ddots & \vdots \\ R_{M,1} & \cdots & R_{M,K} \end{pmatrix},$$

$$\mathbf{\Sigma}_D = \begin{pmatrix} 1 & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & q_{K-1,K} \\ q_{K,1} & \cdots & q_{K,K-1} & 1 \end{pmatrix}. \quad (3)$$

3.2 巨視的変数の微分方程式

式 (1), (2) より，汎化誤差 ϵ_g は R_k と $q_{kk'}$ がすべてわかれば計算できる．自己平均性に基づく一般の学習則の l_k，R_k，$q_{kk'}$ に関する微分方程式は先行研究において以下のように導出されている [6, 7, 8]．ここで t は時刻ステップ n を次元 N で正規化した時刻 $t = n/N$ であり $\langle\cdot\rangle$ はサンプル平均を表す．

$$\frac{dl_k}{dt} = \langle f_k u_k \rangle + \frac{\left\langle f_k^2 \right\rangle}{2l_k}, \quad (4)$$

$$\frac{dR_{mk}}{dt} = \frac{\langle f_k v_m \rangle - \langle f_k u_k \rangle R_{mk}}{l_k} - \frac{R_{mk}}{2l_k^2} \left\langle f_k^2 \right\rangle, \quad (5)$$

$$\frac{dq_{kk'}}{dt} = \frac{\langle f_{k'} u_k \rangle - q_{kk'} \langle f_{k'} u_{k'} \rangle}{l_{k'}} + \frac{\langle f_k u_{k'} \rangle - q_{kk'} \langle f_k u_k \rangle}{l_k} + \frac{\langle f_k f_{k'} \rangle}{l_k l_{k'}} - \frac{q_{kk'}}{2} \left(\frac{\left\langle f_k^2 \right\rangle}{l_k^2} + \frac{\left\langle f_{k'}^2 \right\rangle}{l_{k'}^2} \right). \quad (6)$$

4 結　果

本論文では生徒 $\boldsymbol{J}_k$ の初期値 $\boldsymbol{J}_k^0$，教師 $\boldsymbol{B}_m$ の各要素は平均 0，分散 1 のガウス分布にしたがい独立に生成され，また，$N \to \infty$ の熱力学的極限を考えているので，初期状態においてこれらはすべて直交しており，$R_{mk}^0 = q_{kk'}^0 = 0$ である．このことと教師中間層の対称性，生徒の対称性より，式 (4)–(6) の巨視的変数 $l_k, R_{mk}, q_{kk'}$ から添え字 m, k, k' を落としてそれぞれを l, R, q と書くことにする．

ヘブ学習，パーセプトロン学習，アダトロン学習はそれぞれ以下の式で更新を行う学習則である．

$$f(d,u) = d, \quad (7)$$

$$f(d,u) = \Theta(-ud)\, d, \quad (8)$$

$$f(d,u) = -u\Theta(-ud). \quad (9)$$

$M = 3$ でそれぞれの学習則について式 (4)–(6) を数値的に解いて R,q のダイナミクスを求めた．その際，式 (4)–(6) 中の各サンプル平均はメトロポリス法で求めた．モンテカルロステップ数は 10^6 とした．結果を図 1–3 に示す．

図 1–3 の R, q を使って式 (1) の数値積分を実行することにより汎化誤差 ϵ_g を求めた．数値積分は式 (2) の逆行列の計算を不要にするために積分変数を直交化したうえでメトロポリス法を用いて実行した．その際モンテカルロステップ数は 10^8 とした．結果を図 4–6 に示す．

図 4–6 より，いずれの学習則においても K が大きいほど ϵ_g が小さいことがわかる．すなわち，教師機械がコミティマシンの場合でもアンサンブル学習の効果がある．また，残留汎化誤差はヘブ学習で最も小さく，アダトロン学習で最も大きいことがわかる．図 1–3 よりヘブ学習ではパーセプトロン学習やアダトロン学習よりも q の立ち上がりが速い．すなわち，ヘブ学習では生徒の多様性が急速に失われる．このため，図 4 のようにヘブ学習ではアンサンブルの効果が小さい．図 2 よりパーセプトロン学習では q の定常値が 1 より小さいことがわかる．すなわち，パーセプトロン学習では生徒の多様性が消滅せず残る．そのために図 5 のようにパーセプトロン学習ではアンサンブルの効果が消滅せずに残る．図 3 をよく見るとアダトロン学習では R がいったん最大になり，その後は少し減少することがわかる．また，図 6 よりアダトロン学習では ϵ_g がいったん最小になり，その後少し増大することがわかる．これはヘブ学習やパーセプトロン学習では見られない現象であり興味深い．ϵ_g が最小値をとる時刻は K が大きいほど早く，そのときに生徒アンサンブルの学習モデルが教師モデルにもっとも近くなっていると言える．

謝辞 本論文の一部は科学研究費補助金（課題番号 13780313, 14084212, 14580438, 15500151）によるものである．

参考文献

[1] Freund, Y. and Shapire, R.E.,（安倍直樹訳），人工知能学会誌，**14**(5), 771–780 (1999).

[2] http://www.boosting.org/

[3] 麻生, 津田, 村田, “パターン認識と学習の統計学,” 岩波書店, 東京,2003.

[4] Krogh, A. and Sollich, P., Phys. Rev. E, **55**(1), 811 (1997).

[5] Urbanczik, R., Phys. Rev. E, **62**(1), 1448 (2000).

[6] Miyoshi,S., Hara,K. and Okada,M., cond-mat/0403632

[7] 三好，原，岡田，信学論,**J87-D-II**(7),1391–1401,(2004).

[8] 西森, “スピングラス理論と情報統計力学,” 岩波,1999.

[9] Inoue, J. and Nishimori, H., Phys. Rev. E, **55**(4), 4544 (1997).

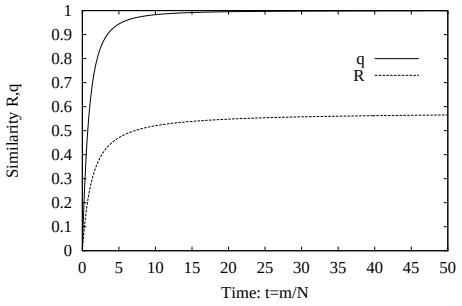

図 1: ヘブ学習の R と q

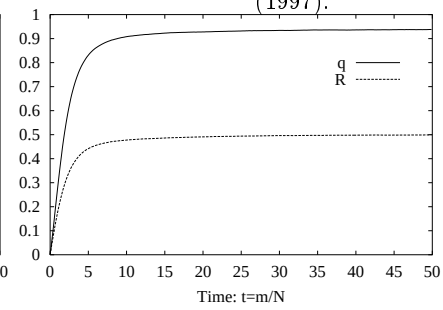

図 2: パーセプトロン学習の R と q

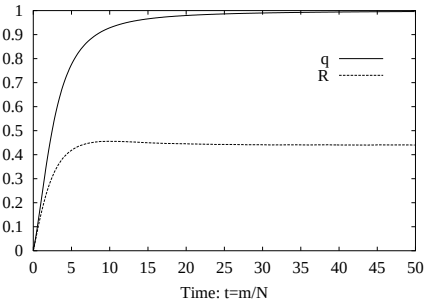

図 3: アダトロン学習の R と q

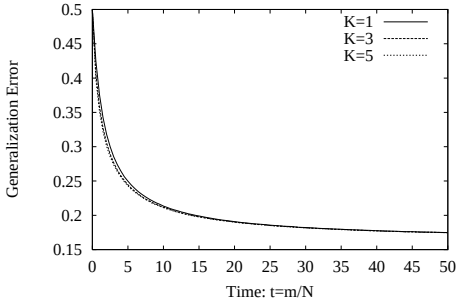

図 4: ヘブ学習の ϵ_g

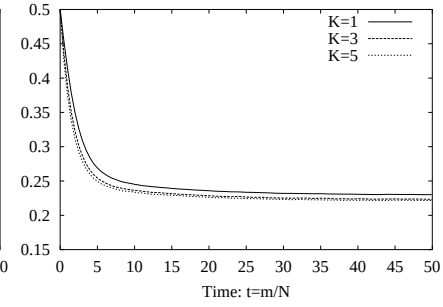

図 5: パーセプトロン学習の ϵ_g

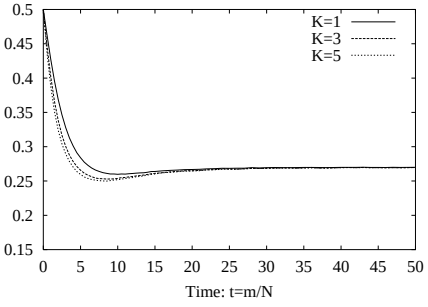

図 6: アダトロン学習の ϵ_g