

# 確率的フィルタリングを用いたアンサンブル学習の統計力学

## Statistical mechanics of ensemble learning using stochastic filtering

三好 誠司 (P)<sup>†</sup>, 岡田 真人<sup>‡</sup>

Seiji Miyoshi(P), Masato Okada

<sup>†</sup>神戸高専 電子工学科 miyoshi@kobe-kosen.ac.jp <sup>‡</sup>東大院 新領域創成科学研究科, 理研 脳総研

**Abstract**— We analyze ensemble learning using Murata's stochastic filtering with statistical mechanical method based on online learning theory. It is shown that the filtering is effective to keep students' variety rich and to improve the generalization performance.

**Keywords**— ensemble learning, online learning, generalization error, stochastic filtering, boosting

### 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うことはアンサンブル学習と呼ばれ注目されている。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 [1]。

アンサンブル学習においては生徒集団の多様性を維持することが重要である。Schapire は生徒の多様性を維持するためにそれぞれの生徒が用いる例題を確率的なフィルタリングにより選択するアルゴリズムを提案した [2]。Murata は Schapire のアルゴリズムをオンライン学習に適した形に修正したアルゴリズムを提案している [3]。このアルゴリズムは生徒の多様性を維持するために、わざと間違った例題(負例)も用いる点が特徴的である。

本稿では Murata のアルゴリズムの効果を理論的に明らかにするため、統計力学的手法を用いた解析を行う。非単調な出力特性を有する教師 [4, 5] を単純パーセプトロンの生徒集団がヘブ則で学習するモデルについて解析を行った結果、フィルタリングの効果により生徒集団の多様性が維持され、汎化能力が大きく改善されることが明らかになった。

### 2 モデル

生徒は 3 個の単純パーセプトロンとし、集団としての出力は生徒の多数決で決定する。各生徒の結合荷重  $J_k$  および入力  $x$  は  $N$  次元ベクトルである。 $J_k$  の初期値  $J_k^0$  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。 $x$  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  $\text{sgn}(u_k l_k)$  である。ここで  $u_k l_k = J_k \cdot x$  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。

教師の出力は  $d = \text{sgn}((v - a)v(v + a))$  である。ここで  $v$  は教師の内部状態であり  $v = B \cdot x$  である。すなわち、教師はしきい値  $a$  の非単調な出力特性を有する [4, 5]。教師の結合荷重  $B$  は  $N$  次元ベクトルであり、その各要素  $B_i$  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。

本稿では  $N \rightarrow \infty$  の熱力学的極限を考えることにする。このとき  $\|x\| = 1$ ,  $\|B\| = \|J_k^0\| = \sqrt{N}$  となる。 $\|J_k\|$  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  $J_k$  の長さと呼ぶことにする。

本稿の記法を用いた場合、Murata の確率的フィルタリングのアルゴリズム [3] は以下の通りである。

**input:** three different classifiers  $J_1, J_2, J_3$

**step1:** pick an example  $(x, d)$

**step2:** choose  $i \in \{1, 2, 3\}$  randomly

**step3:** calculate  $H(x) = \sum_{j \neq i} \text{sgn}(J_j \cdot x)$

**step4:** if  $H(x) = 0$ , train  $J_i$  with  $(x, d)$

otherwise, flip a fair/unfair coin

**head:** train  $J_i$  with  $(x, d)$

**tail:** train  $J_i$  with  $(x, -d)$  負例

**step5:** goto step1 until some condition is fulfilled

このアルゴリズムによる更新は、生徒  $J_1$  を例にとると以下のように書ける。

$$J_1^{m+1} = J_1^m + \Theta(-u_2^m u_3^m) f_1 x^m + \Theta(u_2^m u_3^m) (s^m f_1^m + (1 - s^m) g_1^m) x^m \quad (1)$$

$$P(s^m = 1) = 1 - P(s^m = 0) = b \quad (2)$$

ここで  $\Theta(\cdot)$  はステップ関数である。 $f$  は学習則により決定される関数であり、 $g$  は負例を用いる場合の関数である。たとえばヘブ学習の場合には  $f = -g = d$  となる。また  $b$  は step4 で head となる確率である。

なお、本稿においては統計力学的な取り扱いを容易にするため、step2 は以下とする。

**step2:**  $i = m \pmod{3} + 1$

### 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  $\epsilon_g$  を理論的に求めることである。誤差  $\epsilon$  として、 $\epsilon = \Theta\left(-d \sum_{k=1}^3 \text{sgn}(J_k \cdot x)\right)$  を用いることにする。汎化

誤差  $\epsilon_g$  は  $\epsilon$  を入力  $x$  の確率分布  $p(x)$  で平均したものと定義する。  $\epsilon$  は  $\epsilon(\{u_k\}, v)$  と書くことができるので、  $\epsilon_g$  は確率分布  $p(\{u_k\}, v)$  を用いて、

$$\epsilon_g = \int \left( \prod_{k=1}^3 du_k \right) dv p(\{u_k\}, v) \epsilon(\{u_k\}, v) \quad (3)$$

と書ける。ここで  $p(\{u_k\}, v)$  は  $J_k$  と  $B$  の方向余弦  $R_k$ 、  $J_k$  と  $J_{k'}$  の方向余弦  $q_{kk'}$  を用いて以下のように書ける。

$$p(\{u_k\}, v) = \frac{1}{(2\pi)^2 |\Sigma|^{\frac{1}{2}}} \exp \left( -\frac{(\{u_k\}, v) \Sigma^{-1} (\{u_k\}, v)^T}{2} \right) \quad (4)$$

$$\Sigma = \begin{pmatrix} 1 & q_{12} & q_{13} & R_1 \\ q_{21} & 1 & q_{23} & R_2 \\ q_{31} & q_{32} & 1 & R_3 \\ R_1 & R_2 & R_3 & 1 \end{pmatrix} \quad (5)$$

### 3.2 巨視的変数の微分方程式

式(3)–(5)より、汎化誤差  $\epsilon_g$  は巨視的変数  $R_k, q_{kk'}$  がすべてわかれば計算できる。これらの巨視的変数のダイナミクスを記述する微分方程式は熱力学的極限における自己平均性に基づき式(1)から以下のように決定論的な形で導出できる。ここで  $r_k \equiv l_k R_k$  と  $Q_{kk'} \equiv l_k l_{k'} q_{kk'}$  は解析を容易にするために導入した補助的な巨視的変数である。 $t$  は時刻ステップ  $m$  を次元  $N$  で正規化した時刻  $t = m/N$ 、 $\langle \cdot \rangle$  はサンプル平均を表す。

$$\begin{aligned} \frac{dl_1}{dt} &= \frac{1}{3} \left( \langle \Theta(-u_2 u_3) f_1 u_1 \rangle + b \langle \Theta(u_2 u_3) f_1 u_1 \rangle \right. \\ &\quad \left. + (1-b) \langle \Theta(u_2 u_3) g_1 u_1 \rangle \right) + \frac{1}{6l_1} \left( \langle \Theta(-u_2 u_3) f_1^2 \rangle \right. \\ &\quad \left. + b \langle \Theta(u_2 u_3) f_1^2 \rangle + (1-b) \langle \Theta(u_2 u_3) g_1^2 \rangle \right) \quad (6) \end{aligned}$$

$$\begin{aligned} \frac{dr_1}{dt} &= \frac{1}{3} \left( \langle \Theta(-u_2 u_3) f_1 v \rangle \right. \\ &\quad \left. + b \langle \Theta(u_2 u_3) f_1 v \rangle + (1-b) \langle \Theta(u_2 u_3) g_1 v \rangle \right) \quad (7) \end{aligned}$$

$$\begin{aligned} \frac{dQ_{12}}{dt} &= \frac{1}{3} \left( l_1 \langle \Theta(-u_1 u_3) f_2 u_1 \rangle + b l_1 \langle \Theta(u_1 u_3) f_2 u_1 \rangle \right. \\ &\quad \left. + (1-b) l_1 \langle \Theta(u_1 u_3) g_2 u_1 \rangle + l_2 \langle \Theta(-u_2 u_3) f_1 u_2 \rangle \right. \\ &\quad \left. + b l_2 \langle \Theta(u_2 u_3) f_1 u_2 \rangle + (1-b) l_2 \langle \Theta(u_2 u_3) g_1 u_2 \rangle \right) \quad (8) \end{aligned}$$

## 4 結果と考察

理論的に求めた  $\epsilon_g$  を計算機実験 ( $N = 5000$ ) の結果と重ねて図1に示す。生徒の学習則はヘブ則とした。理論計算においては生徒の対称性を仮定し、非単調教師のしきい値は  $a = 0.6$  とした。step4でheadとなる確率は  $b = 0.2, 0.5, 0.8$  の3通りについて計算した。式(6)–(8)に現れるサンプル平均をモンテカルロステップ数  $MC = 10^6$  のメトロポリス法で求めながら、ルンゲクッタ法により連立微分方程式を数値的に解いた。得られた  $R, q$  を用いて式(3)の積分を再度メトロポリス法 ( $MC = 10^9$ ) で実行することにより  $\epsilon_g$  を求めた。

図中の“Normal”は通常のアンサンブル学習の結果であり、この場合、 $\epsilon_g$  は非単調なダイナミクスを呈する[5]。通常のアンサンブル学習と比較すると、学習の初期を除いては、フィルタリングにより汎化能力が大きく改善されていることがわかる。

このときの  $R$  と  $q$  の関係を図2に示す。いずれの曲線も  $R = q = 0$  の端点が  $t = 0$  に対応し、もう一方の

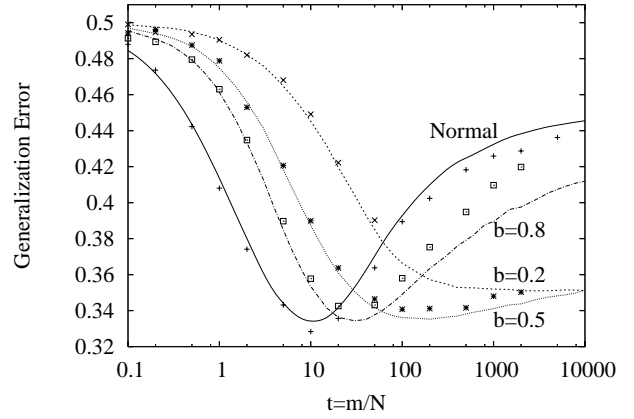


図1: 汎化誤差  $\epsilon_g$  のダイナミクス ( $a = 0.6$ )

端点が  $t = 10000$  に対応する。フィルタリングにより、 $R$  に対して  $q$  が非常に小さく抑えられており、生徒集団の多様性が維持されていることがわかる。図1より  $b = 0.2$  と  $b = 0.5$  は  $t = 10000$  における  $\epsilon_g$  がほぼ等しい値になっているが、図2より、その値は全く異なる  $R, q$  によって実現されていることがわかり興味深い。

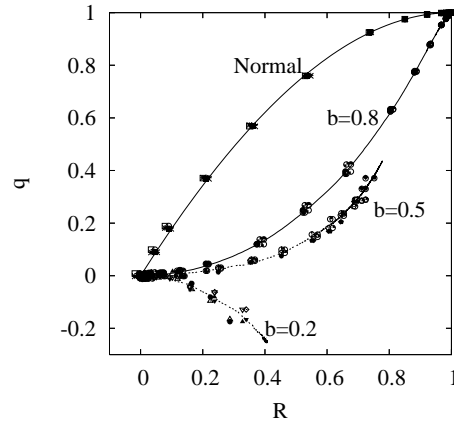


図2:  $R$  と  $q$  の関係 ( $a = 0.6$ )

謝辞 本研究の一部は科学研究費補助金(課題番号14084212, 14580438, 15500151)によるものである。

### 参考文献

- [1] S. Miyoshi, K. Hara and M. Okada(2005) “Analysis of ensemble learning using simple perceptrons based on online learning theory.”, Phys. Rev. E, **71**: 036116.
- [2] R.E. Schapire(1990) “The strength of weak learnability.”, Machine Learning, **5**: 197–227.
- [3] N. Murata(2005) “Stochastic filtering for on-line boosting.”, NHC & SMAPIP-2005, Proc. 2005 Workshop on Randomness and Computation: 41–42.
- [4] J.I. Inoue and H. Nishimori(1997) “On-line AdaTron learning of a unlearnable rules.”, Phys. Rev. E, **55**(4): 4544–4551.
- [5] 三好, 原, 岡田 (2005) “教師が非単調な場合のアンサンブル学習”, 信学技報, NC2004-214: 123–128.