アンサンブル学習モデルにおける重み付けに関する考察

内田　真人†（正員）　　塩谷　浩之††（正員）

A Study on Assignment of Weight Parameters in Ensemble Learning Model

Masato UCHIDA† and Hiroyuki SHIOYA††, *Regular Members*

† 日本電信電話株式会社　NTT サービスインテグレーション基盤研究所，武蔵野市
NTT Service Integration Laboratories, NTT Corporation, 3–9–11 Midori-cho, Musashino-shi, 180–8585 Japan

†† 室蘭工業大学工学部情報工学科，室蘭市
Department of Computer Science and Systems Engineering, Muroran Institute of Technology, 27–1 Mizumoto-cho, Muroran-shi, 050–8585 Japan

あらまし　アンサンブル学習とは，個々に学習された複数の学習機械の予測結果の重み付け平均を最終的な予測結果として用いる手法である．この学習法は，Kullback divergence に関する 3 段階の最小化操作として理解できる場合がある．本論文では，この枠組みにおける重み付けに関する考察を行う．

キーワード　アンサンブル学習，重み付け，初期値変動法，三角不等式

1. ま え が き

個々に学習された複数の学習機械が与えられたとき，汎化能力の高い予測機械をどのように得るかという問題は重要である．アンサンブル学習とは，与えられた機械の予測結果の重み付け平均を最終的な予測結果とすることでこの問題への対応を試みる手法である．この種の学習法の手続きは様々に設定可能であると思われるが，学習機械の混合に関する確率モデルを適当に設定することで，学習法の手続きを明確に表現できる場合がある [4]．この概要を以下に示す．

集合 $\boldsymbol{\mathcal{Z}}$ $(\subset \mathbb{R}^d)$ 上の確率分布全体を

$$\mathcal{P}(\boldsymbol{\mathcal{Z}}) \stackrel{\text{def}}{=} \left\{ p \,\middle|\, p : \boldsymbol{\mathcal{Z}} \to \mathbb{R}, p(\boldsymbol{z}) > 0 \ (\forall \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}), \int_{\boldsymbol{\mathcal{Z}}} p(\boldsymbol{z}) d\boldsymbol{z} = 1 \right\}$$

とおく．ここで，$p_i(\boldsymbol{z})$ $(\in \mathcal{P}_i(\boldsymbol{\mathcal{Z}}) \subset \mathcal{P}(\boldsymbol{\mathcal{Z}}))$ を用いて $(i = 1, \cdots, M)$，新たな確率分布 $\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})$ $(\in \mathcal{P}(\boldsymbol{\mathcal{Z}}))$ を

$$\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{\prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i}}{\int_{\boldsymbol{\mathcal{Z}}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z}} \tag{1}$$

と定義する．ただし

$$\int_{\boldsymbol{\mathcal{Z}}} \prod_{i=1}^{M} p_i(\boldsymbol{z})^{\beta_i} d\boldsymbol{z} < \infty \tag{2}$$

$$\sum_{i=1}^{M} \beta_i = 1, \quad \beta_i > 0 \tag{3}$$

$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)^T \in \mathbb{R}^M \tag{4}$$

とする．

このとき，以下のような Kullback divergence に関する 3 段階の最小化操作を考える．これは，2 乗誤差関数を損失関数とし，与えられた機械の出力の重み付け平均を予測値とするという学習法 [3] の自然な一般化である [4]．ただし，$D(\cdot\|\cdot)$ は Kullback divergence を表し，p_* は推定の対象となる真の分布である．

$$\hat{p}_i(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg \min_{p(\boldsymbol{z}) \in \mathcal{P}_i(\boldsymbol{\mathcal{Z}})} D(p_* \| p) \tag{5}$$

$$\bar{\hat{p}}_{\boldsymbol{\beta}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \arg \min_{p(\boldsymbol{z}) \in \mathcal{P}(\boldsymbol{\mathcal{Z}})} \sum_{i=1}^{M} \beta_i D(p \| \hat{p}_i) \tag{6}$$

$$\hat{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^M} D(p_* \| \bar{\hat{p}}_{\boldsymbol{\beta}}) \tag{7}$$

上記の手続きにより求められた $\bar{\hat{p}}_{\hat{\boldsymbol{\beta}}}$ が，最終的に推定された確率分布であり，各 $\hat{p}_i$ $(i = 1, \cdots, M)$ が事前に推定された確率分布である．この枠組みを用いることにより，与えられたサンプルを分割し個々の学習機械を生成するという手法に基づくアンサンブル学習は，平均予測誤差の意味で有効ではないという事実が導かれる（ [4] の系 1）．

一方で，[4] においては $\bar{p}_{\boldsymbol{\beta}}$ のパラメータ $\boldsymbol{\beta}$ に関する議論が十分になされていない．例えば，$\bar{p}_{\boldsymbol{\beta}}$ が確率分布となるためには式 (2) の条件が与えられていれば十分であるのにもかかわらず，この条件に加えて，パラメータ $\boldsymbol{\beta}$ が重みとしての意味をもつための条件である式 (3) が追加されている点に関する考察が十分ではない．更に，式 (7) により定まる $\hat{\boldsymbol{\beta}}$ がこの条件を満たすかどうかも全く自明ではない．

次章では，与えられたサンプルをすべて用い個々の学習機械を生成する，という手法に基づくアンサンブル学習の性質を重み $\boldsymbol{\beta}$ の最適化の観点から述べ，更に，式 (3) がアンサンブル学習の結果とどのようにかかわってくるのかを調べる．

2. 主　結　果

式 (7) により定まる $\hat{\boldsymbol{\beta}}$ の特徴は，式 (1) により定められる確率モデルが指数型分布族に属することから

$$D(p\|\bar{p}_{\hat{\boldsymbol{\beta}}}) = D(p\|p_i) - D(\bar{p}_{\hat{\boldsymbol{\beta}}}\|p_i), \quad (i = 1, \cdots, M) \tag{8}$$

のように与えられることが知られている [1].

式 (8) は，アンサンブル学習において，与えられたすべてのサンプルを用いた上で，パラメータをこう配法により決定する際の初期値を変えて得られた複数の学習機械を利用すること（初期値変動法 [3]）が，[4] の系 1 によって得られた，アンサンブル学習に関する否定的な結果を解決する一つの方法であることを示している．なぜならば，$\mathcal{Q}(\boldsymbol{\mathcal{Z}}) = \{p_1(\boldsymbol{z}), \cdots, p_M(\boldsymbol{z})\}$ を，異なる初期値を用いて学習した M 個の学習機械とし，一般性を失うことなく

$$p_M(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \mathcal{Q}(\boldsymbol{\mathcal{Z}})} D(p\|q) \tag{9}$$

としたとき，最適な重み $\hat{\boldsymbol{\beta}}$ が得られたとすれば，式 (8) より，$D(p\|\bar{p}_{\hat{\boldsymbol{\beta}}})$ は $D(p\|p_M)$ よりも $D(\bar{p}_{\boldsymbol{\beta}}\|p_M)$ だけ小さくなるからである．この様子は [3] において実験的に示されていた．

［注意 ］ サンプルを分割するという手法においては，$\boldsymbol{\beta}$ の最適化を行っても，上記のような性質は得られない（ [4] の系 1 参照）． □

ところで，このときの $\hat{\boldsymbol{\beta}}$ は式 (3) の条件を必ずしも満たさない．つまり，$\hat{\boldsymbol{\beta}}$ を用いた場合には，重み付き平均をとるという意味を与えられない場合があるということである．しかし，次が成り立つ．

［定理 ］ 式 (8) における $\hat{\boldsymbol{\beta}}$ が一意に定まるとき

$$0 \le \hat{\beta}_i \le 1, \quad (i = 1, \cdots, M) \tag{10}$$

が成り立つための必要十分条件は

$$D(p\|p_M) + D(p_M\|p_i) \ge D(p\|p_i)$$

である $(i = 1, \cdots, M)$．ただし，一般性を失うことなく

$$D(p\|p_M) = \min_{i=1,\cdots,M} D(p\|p_i)$$

と仮定した． □

（証明については付録を参照）

上記の定理から，式 (10) を満たすためには，各 i について，p_i 及び p_M の p（真の分布）に対する近似精度の差 $D(p\|p_i) - D(p\|p_M)$ が $D(p_M\|p_i)$ より小さいことが要請される．これは，p_i の近似精度が悪い場合には，式 (10) が成り立たないことを意味する．

なお，式 (10) が成り立たない場合，ある j, k について $\beta_k > 1, \beta_j < 0$ が成り立つ．これは，$\beta_j (< 0)$ により p_j の高密度部を低密度部に変換し，更に，$\beta_k (> 0)$ により p_k の高密度部を際立たせることになる．無論，上記の p_k は複数の場合もある．

3. むすび

本論文では，アンサンブル学習における重み $\boldsymbol{\beta}$ の最適化に関する考察を加えた．その結果，最適な重み $\hat{\boldsymbol{\beta}}$ が得られたとすれば，[4] の系 1 により結論されるアンサンブル学習に関する否定的な結果を解決する一つの方法として，初期値変動法は有効であるということがわかった．この結果や [4] の系 1 の主張は，本論文で用いたアンサンブル学習の理論モデルの特徴に基づき導出されたものである．したがって，式 (5)，(6)，(7) により与えられる本論文の枠組みは，[3] の数値実験により示されている結果を説明するために適した理論モデルの一つであるといえる．

一方，$\hat{\boldsymbol{\beta}}$ については，それがアンサンブル学習における重みとしての意味をもつための必要十分条件を Kullback divergence に関する三角不等式により与えた．

今後の課題としては，近似精度の低い学習機械を用いた場合のアンサンブル学習における汎化誤差解析などが挙げられる．

文　献

[1] 甘利俊一，長岡浩司，情報幾何の方法，岩波講座 応用数学 6 [対象 12]，岩波書店，1993.

[2] 大関信雄，青柳雅計，不等式，槙書店，1976.

[3] 上田修功，中野良平 “アンサンブル学習における汎化誤差解析，” 信学論（D-II），vol.J80-D-II, no.9, pp.2512–2521, Sept. 1997.

[4] 内田真人，塩谷浩之，伊達　惇 “アンサンブル学習の解析と拡張，” 信学論（D-II），vol.J84-D-II, no.7, pp.1537–1542, July 2001.

付　録

定理の証明

一般性を失うことなく

$$\beta_M = 1 - \sum_{i=1}^{M-1} \beta_i$$

としてよい．このとき

$$\frac{\partial}{\partial \beta_j} D(p\|\bar{p}_{\boldsymbol{\beta}}) = D(p\|p_j) - D(p\|p_M) + \int_{\boldsymbol{\mathcal{Z}}} \bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}) \log \frac{p_j(\boldsymbol{z})}{p_M(\boldsymbol{z})} d\boldsymbol{z} \quad \text{(A·1)}$$

が成り立つ．よって

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta_j^2}D(p\|\bar{p}_{\boldsymbol{\beta}})\\
&\quad=\int_{\boldsymbol{\mathcal{Z}}}\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})\left\{\log\frac{p_j(\boldsymbol{z})}{p_M(\boldsymbol{z})}\right\}^2 d\boldsymbol{z}\\
&\qquad-\left\{\int_{\boldsymbol{\mathcal{Z}}}\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})\log\frac{p_j(\boldsymbol{z})}{p_M(\boldsymbol{z})}d\boldsymbol{z}\right\}^2\\
&\quad=\int_{\boldsymbol{\mathcal{Z}}}\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z})\left\{\log\frac{p_j(\boldsymbol{z})}{p_M(\boldsymbol{z})}\right.\\
&\qquad\left.-\int_{\boldsymbol{\mathcal{Z}}}\bar{p}_{\boldsymbol{\beta}}(\boldsymbol{z}')\log\frac{p_j(\boldsymbol{z}')}{p_M(\boldsymbol{z}')}d\boldsymbol{z}'\right\}^2 d\boldsymbol{z}\\
&\quad\geq 0
\end{aligned}$$

となるので，$D(p\|\bar{p}_{\boldsymbol{\beta}})$ は β_j に関して下に凸である $(j=1,\cdots,M-1)$．また，式 (A·1) より

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}D(p\|\bar{p}_{\boldsymbol{\beta}})\bigg|_{\beta_j=1}&=D(p\|p_j)-D(p\|p_M)\\
&\quad+D(p_j\|p_M)\geq 0
\end{aligned}$$

が成り立つ $(j=1,\cdots,M-1)$．したがって

$$0\leq\hat{\beta}_j\leq 1$$

であるための必要十分条件は

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}D(p\|\bar{p}_{\boldsymbol{\beta}})\bigg|_{\beta_j=0}&=D(p\|p_j)-D(p\|p_M)\\
&\quad-D(p_M\|p_j)\leq 0
\end{aligned}$$

である $(j=1,\cdots,M-1)$．この不等式は $j=M$ のときには自明に成り立つ．一方

$$\begin{aligned}
&0\leq\hat{\beta}_j\leq 1,\quad(j=1,\cdots,M-1)\\
&\quad\rightarrow\hat{\beta}_M=1-\sum_{i=1}^{M-1}\hat{\beta}_i\leq 1
\end{aligned}$$

が成り立つ．よって

$$\begin{aligned}
&0\leq\hat{\beta}_j\leq 1,\quad(j=1,\cdots,M-1)\\
&\quad\rightarrow\hat{\beta}_M=1-\sum_{i=1}^{M-1}\hat{\beta}_i\geq 0
\end{aligned}$$

が成り立てば証明は完了する．これを背理法で示すために $\hat{\beta}_M<0$ と仮定する．このとき，一般性を失うことなく $\beta_j\neq 0$ $(j=1,\cdots,M-1)$ とすると

$$\begin{aligned}
&\int_{\boldsymbol{\mathcal{Z}}}\prod_{i=1}^{M}p_i(\boldsymbol{z})^{\hat{\beta}_i}d\boldsymbol{z}\\
&\quad\geq(1-\hat{\beta}_M)\int_{\boldsymbol{\mathcal{Z}}}\left\{\prod_{i=1}^{M-1}p_i(\boldsymbol{z})^{\hat{\beta}_i}\right\}^{\frac{1}{1-\hat{\beta}_M}}d\boldsymbol{z}\\
&\qquad+\hat{\beta}_M\int_{\boldsymbol{\mathcal{Z}}}p_M(\boldsymbol{z})d\boldsymbol{z}\\
&\quad\geq\int_{\boldsymbol{\mathcal{Z}}}\left\{\prod_{i=1}^{M-1}p_i(\boldsymbol{z})^{\hat{\beta}_i}\right\}^{\frac{1}{1-\hat{\beta}_M}}d\boldsymbol{z}\\
&\qquad+\hat{\beta}_M\int_{\boldsymbol{\mathcal{Z}}}\left\{p_M(\boldsymbol{z})-\sum_{i=1}^{M-1}\frac{\hat{\beta}_i}{1-\hat{\beta}_M}p_i(\boldsymbol{z})\right\}d\boldsymbol{z}\\
&\quad=\int_{\boldsymbol{\mathcal{Z}}}\prod_{i=1}^{M-1}p_i(\boldsymbol{z})^{\frac{\hat{\beta}_i}{1-\hat{\beta}_M}}d\boldsymbol{z}
\end{aligned}$$

が成り立つ．ここで，$a>0$，$b>0$，$\frac{1}{u}+\frac{1}{v}=1$，$(u\neq 0)$ のとき

$$\begin{aligned}
&u<1\quad\text{ならば}\quad a^{\frac{1}{u}}b^{\frac{1}{v}}\geq\frac{a}{u}+\frac{b}{v}\\
&u>1\quad\text{ならば}\quad a^{\frac{1}{u}}b^{\frac{1}{v}}\leq\frac{a}{u}+\frac{b}{v}
\end{aligned}$$

であり，等号は $a=b$ のときかつそのときに限る [2] ことを用いた．また

$$\begin{aligned}
&\sum_{i=1}^{M}\hat{\beta}_iD(p\|p_i)\\
&\quad=\sum_{i=1}^{M-1}\frac{\hat{\beta}_i}{1-\hat{\beta}_M}D(p\|p_i)+\hat{\beta}_MD(p\|p_M)\\
&\qquad+\sum_{i=1}^{M-1}\left(\hat{\beta}_i-\frac{\hat{\beta}_i}{1-\hat{\beta}_M}\right)D(p\|p_i)\\
&\quad\geq\sum_{i=1}^{M-1}\frac{\hat{\beta}_i}{1-\hat{\beta}_M}D(p\|p_i)+\hat{\beta}_MD(p\|p_M)\\
&\qquad+\sum_{i=1}^{M-1}\left(\hat{\beta}_i-\frac{\hat{\beta}_i}{1-\hat{\beta}_M}\right)D(p\|p_M)\\
&\quad=\sum_{i=1}^{M-1}\frac{\hat{\beta}_i}{1-\hat{\beta}_M}D(p\|p_i)
\end{aligned}$$

が成り立つ．よって

$$\begin{aligned}
&\tilde{\beta}_j=\frac{\hat{\beta}_j}{1-\hat{\beta}_M},\quad j=1,\cdots,M-1\\
&\tilde{\boldsymbol{\beta}}=(\tilde{\beta}_1,\cdots,\tilde{\beta}_{M-1})^T
\end{aligned}$$

とおけば，[4] の補題 1 と $\hat{\boldsymbol{\beta}}$ の一意性より

$D(p\|\bar{p}_{\tilde{\boldsymbol{\beta}}}) < D(p\|\bar{p}_{\hat{\boldsymbol{\beta}}})$

が成り立つ．これは，$\hat{\boldsymbol{\beta}}$ の定義に矛盾する．すなわち，$\hat{\beta}_M \geq 0$ である．

（平成 14 年 7 月 10 日受付，12 月 19 日再受付）