社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

カルバック情報量の分割による特異点モデルの学習係数計算法

永田　賢二†　　渡辺　澄夫††

† 東京工業大学 工学部 情報工学科 〒 226-8503 横浜市緑区長津田 4259
†† 東京工業大学 精密工学研究所 〒 226-8503 横浜市緑区長津田 4259
E-mail: †kenji.nagata@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

あらまし　神経回路網、混合正規分布、ベイズネットワーク、隠れマルコフモデル、ボルツマンマシンなど、近年情報科学において広く用いられるようになった学習モデルの多くは特異モデルであることが知られている。これらのフィッシャー情報行列は正定値ではないので、従来の統計的漸近理論が成り立たず、学習モデル選択アルゴリズムはまだかくりつされていない。本論ではカルバック情報量を分割することにより、学習係数を計算することを提案し、その有効性をいくつかの実験により明らかにする。
キーワード　特異モデル、カルバック情報量、確率的複雑さ

A Method to Estimate the Learning Coefficients of Singular Learning Machines by Decomposition of Kullback Information

Kenji NAGATA† and Sumio WATANABE††

† Department of Computer Science Tokyo Institute of Technology, 4259
Nagatsuda,Midori-ku,Yokohama,226-8503 Japan
†† PI Lab., Tokyo Institute of Technology.
E-mail: †kenji.nagata@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

Abstract　A lot of learning machines such as neural networks, normal mixtures, Bayesian networks, and hidden Markov models are singular statistical models. Their Fisher information matrices are not positive definite, hence the conventional statistical asymptotic theory does not hold. In this paper, we propose a new method to calculate the learning coefficients by decomposing the Kullback information. The effectiveness of the proposed method is shown by experimetal results.
Key words　Singular Learning Machines , Kullback Information , Stochastic Complexity

1. まえがき

神経回路網、混合正規分布、ベイズネットワーク、隠れマルコフモデル、ボルツマンマシンなど、例からの学習によって確率的な推論方式を獲得するモデルは、階層的な構造や対称性に起因するフィッシー情報行列の縮退が生じるため、統計的正則モデルではないことが知られている。これらの学習モデルは総称して特異モデルと呼ばれているが、特異モデルにおいては、最尤推定量の分布もベイズ事後分布も正規分布に漸近しないために、従来の統計学的あるいは統計物理学的な理論は成り立たず、最適な設計法が未だに確立されていない [1] [6] [7] [8] [9] [10] [11]。

学習理論においては、「例からの学習」についての数学的な理論と情報学的なアルゴリズムが研究されているが、その主要な目的は次のふたつである。

(1) 順問題：サンプルを発生している真の確率分布 $q(x)$、真の分布を推測するために用いられる学習モデル $p(x|w)$ およびその事前分布 $\varphi(w)$ が与えられたとき、「学習結果と真の分布との違い」を解明すること。

(2) 逆問題：真の分布 $q(x)$ が不明であって、サンプルだけが与えられたときに、「学習結果と真の分布との違い」をできるだけ小さくするように、学習モデルや事前分布を定めるアルゴリ

ズムを作ること。

このうち、情報科学への応用においては、(2) の逆問題が重要であるが、逆問題を考えるためには、そのための基礎となる (1) の順問題を解決しておく必要がある。

統計的正則モデルにおいては、最尤推定量の漸近分布やベイズ事後分布の漸近形を容易に導くことができ、順問題について解明されているので、その結果に基づいて、逆問題へのアプローチとして AIC,BIC,MDL などの学習モデル選択のアルゴリズムが考案されている。

しかしながら、特異モデルにおいては、未だに順問題が解明されていないために、数学的に意味のある学習アルゴリズムはまだ確立されていない。そこで、本論では、特異モデルの順問題について、カルバック情報量を分割することにより、確率的複雑さを分解する方法を提案し、その有効性をいくつかの実験により明らかにする。

2. 特異モデルの学習理論

2.1 ベイズ学習の枠組

本論では、特異モデルの学習において、現在のところ、最も高精度な推測を実現すると考えられているベイズ法について検討する。

N 次元ユークリッド空間 R^N 上の確率変数 X が確率密度関数 $q(x)$ を持つとし、X の独立なサンプルが n 個得られた場合を考える。そのサンプルを

$$X^n = (X_1, X_2, ..., X_n)$$

と書く。サンプルは、それが観測される度に確率的に変動するのであるから、確率変数である。このサンプルから真の分布を推測するために d 次元ユークリッド空間 R^d 内に値を取るパラメータ w を持つ確率推論モデル $p(x|w)$ を考える。$p(x|w)$ は、神経回路網、混合正規分布、ベイズネットワークなどの高度に複雑な学習モデルを表すものとする。またパラメータ w が従う確率分布 (事前分布) を $\varphi(w)$ とする。このとき、サンプル X^n とパラメータ w の同時分布は

$$p(X^n, w) = \varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

であるから、「サンプル X^n が与えられたという条件下におけるパラメータ w の分布」は、ベイズの定理によって

$$p(w|X^n) = \frac{1}{Z(X^n)} \varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

となる。ここで、$Z(X^n)$ は正規化定数

$$Z(X^n) = \int dw \varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

であるが、これは、学習モデルと事前分布 $(p(x|w), \varphi(w))$ の尤度に等しい (学習モデルと事前分布の「証拠」と呼ばれることもあり、統計物理学における分配関数と同じである)。ベイズ学習においては、この事後分布で学習モデルを平均することにより、予測分布 $p(x|X^n)$ を構成する。すなわち、

$$p(x|X^n) = \int p(x|w) p(w|X^n) dw.$$

この予測分布 $p(x|X^n)$ は、「サンプル X^n が与えられたというもとで X の密度関数を推測したもの」である。予測分布は真の分布に近いと思われるが、サンプルが有限であることおよびサンプルがゆらぎを持つことから、完全には一致しない。学習理論における最初の課題は、「予測分布と真の分布 $q(x)$ の違いは、サンプルが増えるにつれて、どのような早さで小さくなってゆくか」を明らかにすることである。

2.2 カルバック情報量

ここで、一般に二つの確率分布 $q(x)$ と $p(x)$ の違いを表す量としてカルバック情報量を導入する。それは

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

によって定義される。一般に $D(q||p) \neq D(p||q)$ なので、カルバック情報量は、通常の意味での距離にはならないが、任意の $q(x), p(x)$ に対して

$$D(q||p) \geqq 0$$

が成り立ち、また

$$D(q||p) = 0 \Longrightarrow q(x) = p(x) \ \ (^\forall x)$$

が成り立つ。

さて、学習理論においては、次のふたつのカルバック情報量が重要である。ひとつは、真の分布と学習モデルとのカルバック情報量である。

$$H(w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx$$

もうひとつは、真の分布とベイズ予測分布とのカルバック情報量である。

$$G(X^n) = \int q(x) \frac{q(x)}{p(x|X^m)} dx$$

ある。$H(w)$ は真の分布と学習モデルが与えられれば、パラメータの関数として確定する。一方、$G(X^n)$ は、サンプル X^n の出方によって確率的に変動する確率変数である。$G(X^n)$ は「ベイズ予測分布が、どの程度真の分布に近いか」を表すので、汎化誤差と呼ばれている。カルバック情報量 $H(w)$ に基づいて、汎化誤差がどのような性質をもっているかを解明することが、学習理論における順問題である。

2.3 特異モデルの性質

特異モデルの学習理論によって、汎化誤差 $G(X^n)$ は、ある定数 λ が存在して、次のように表されることが知られている [10]。

$$E[G(X^n)] = \frac{\lambda}{n} + o(\frac{1}{n}).$$

ここで λ は学習係数と呼ばれる。学習における順問題は、学習係数 λ を求めることに帰着するが、これまでの研究によって、学習係数は、次の幾つかの特徴づけができることが知られている。

(1) $(-\lambda)$ は、ゼータ関数

$$\zeta(z) = \int H(w)^z \varphi(w) dw$$

の最も原点に近い極に等しい [10]。

(2) λ は次の極限値である [9]。

$$\lambda = \lim_{n\to\infty} \frac{-\log \int \exp(-nH(w))\varphi(w)dw}{\log n}.$$

(3) λ は次の極限値である [11]。

$$\lambda = \lim_{t\to 0} \frac{\log(V(at)/V(t))}{\log t}$$

ここで

$$V(t) = \int_{H(w)<t} \varphi(w) dw.$$

このうち、(1) は、カルバック情報量の特異点を解消することにより学習係数が求められることを述べており、様々な学習モデルの学習係数を理論的に求める場合に用いられる。また (2)(3) はカルバック情報量と事前分布が与えられた場合に、学習係数を数値計算で求める場合に利用される。

3. 提 案 方 法

本論では、カルバック情報量が

$$H(w) = H_1(w) + H_2(w)$$

と分解されて、かつ $H_1(w)$ については学習係数 λ_1 が解明されている場合に、$H(w)$ の学習係数 λ を求めるアルゴリズムを提案し、その有効性を実験的に明らかにする。まず、提案するアルゴリズムの先だって、その基礎となる定理を述べて証明を行う。

3.1 基礎定理と証明

確率分布 $\rho_n(w)$ を次のように定義する。

$$\rho_n(w) \propto \exp(-nH_1(w))\varphi(w).$$

このとき、次の定理がなりたつ。

定理 1 学習係数について次の関係が成立つ。

$$\lambda = \lambda_1 + \lim_{n\to\infty} \frac{-\log \int \exp(-nH_2(w))\rho_n(w)dw}{\log n}$$

(定理 1 の証明)

$$Z(n) = \int e^{-nH(w)}\varphi(w)dw$$

とおくと、学習係数の定義から、

$$\lambda = \lim_{n\to\infty} \frac{-\log Z(n)}{\log n}$$

であるが、これに関係式

$$Z(n) = \int e^{-nH_1(w)}\varphi(w)dw \times \frac{\int e^{-nH_2(w)-nH_1(w)}\varphi(w)dw}{\int e^{-nH_1(w)}\varphi(w)dw}$$

を代入すると定理が得られる（定理 1 の証明終わり）。

3.2 計算アルゴリズム

上記の定理より、学習係数 λ を数値的に求めるアルゴリズムとして次の方式を作ることができる。

計算アルゴリズム

1. n として幾つかの値を設定する。

2. $\rho_n(w)$ に従うサンプル $\{w_k; k = 1, 2, ..., K\}$ を取り出す。

3.

$$y(n) = -\log\{\frac{1}{K}\sum_{k=1}^{K} \exp(-nH_2(w_k))\}$$

を計算する。

4. 幾つかの n についてくみあわせ $(\log n, y(n))$ を求めて、回帰曲線 $y = ax + b$ を最小二乗法で当てはめることにより a^* を求める。

5. $\lambda_1 + a$ を目的の値とする。

このアルゴリズムにおいて、２．の手続きにおいては、次のメトロポリス法を用いることにする。

メトロポリス法

$\rho_n(w)$ に従う確率分布に法則収束する $\{w_k\}$ は次のようにして作り出せる。

１．w の初期値を設定する。

２．w に乱数を加えることにより w' を得る。

３．もしも $H_1(w) > H_1(w')$ であれば、w' を採択し、１．の手続きにもどる。

４．もしも $H_1(w) \leqq H_1(w')$ であれば、確率

$$P = \exp(-nH_1(w') + nH_1(w))$$

で w' を採択し、確率 $1-P$ で w を採択し、１．の手続きに戻る。

4. 実　　験

4.1 実験の条件

提案するアルゴリズムの評価を行うため、次の３ケースについて実験を行った。本論で提案した方法は、学習係数が未知であるカルバック情報量に適用することが目的であるが、以下では、提案する方法の有効性を確認するために学習係数が理論的に解明されているモデルを調べることにする。

(1) 入力１個、中間ユニット２個、出力ユニット１個の三層パーセプトロンのカルバック情報量は

$$\begin{aligned} H(a,b,c,d) &= H_1(a,b,c,d) + H_2(a,b,c,d) \\ H_1(a,b,c,d) &= (ab+cd)^2 \\ H_2(a,b,c,d) &= (ab^3+cd^3)^2 \end{aligned}$$

となることが知られている。このモデルでは、$H(a,b,c,d)$ および $H_1(a,b,c,d)$ ともに特異点解消を求めることができて、$\lambda = 2/3,\ \lambda_1 = 1/2$ になることが知られている。そこでこの場合を用いることにすると $a^* = 1/6$ である。

(2) 入力１個、中間ユニット２個、出力ユニット１個の三層パーセプトロンのカルバック情報量を次のように分割する。

$$\begin{aligned} H(a,b,c,d) &= H_1(a,b,c,d) + H_2(a,b,c,d) \\ H_1(a,b,c,d) &= (ab+cd)^2 + (1/n)(ab^3+cd^3)^2 \\ H_2(a,b,c,d) &= (1-1/n)(ab^3+cd^3)^2 \end{aligned}$$

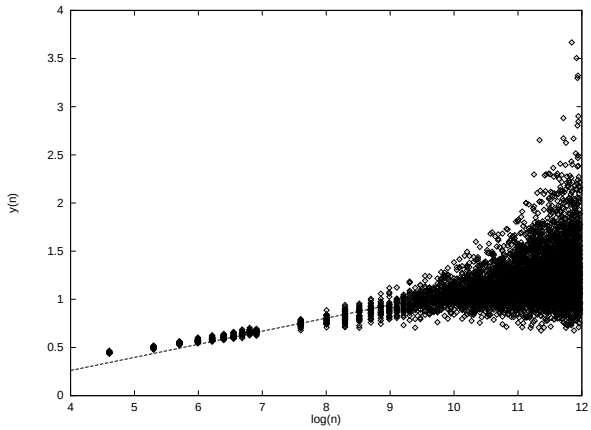

図１　実験１の結果

この分割において、$1/n$ は $n \to \infty$ において０に収束するので、$\lambda = 2/3,\ \lambda_1 = 1/2$ になることは代わらない。そこでこの場合を用いることにすると $a^* = 1/6$ である。

(3) 入力１個、中間ユニット３個、出力ユニット１個の三層パーセプトロンのカルバック情報量は

$$\begin{aligned} H(a,b,c,d,e,f) &= H_1(a,b,c,d,e,f) \\ &\quad + H_2(a,b,c,d,e,f) \\ H_1(a,b,c,d,e,f) &= (ab+cd+ef)^2 \\ H_2(a,b,c,d,e,f) &= (ab^3+cd^3+ef^3)^2 \\ &\quad + (ab^5+cd^5+ef^5)^2 \end{aligned}$$

このモデルについても、特異点解消によって学習係数は得られていて、$\lambda = 5/6,\ \lambda_1 = 1/2$ である。そこでこの場合には $a^* = 1/3$ である。

注意 入力１個、中間ユニット H 個、出力ユニット１個の３層パーセプトロンについては、帰納的な特異点解消によって、λ の値が理論的に求められている [2]。K を $K^2 \leqq H$ を満たす最大の自然数として、

$$\lambda = \frac{H + K^2 + K}{4K+2}$$

が成り立つ。

4.2 実 験 結 果

MCMC 法は、実験を行なうたびに確率的に変動する結果になるので、ひとつの n について 30 通りの実験を行なった。上記で述べた３通りのケースについて、最小二乗法で直線を推測した結果を図１、図２、図３に示す。

(1) 最小二乗法で得られた回帰曲線は

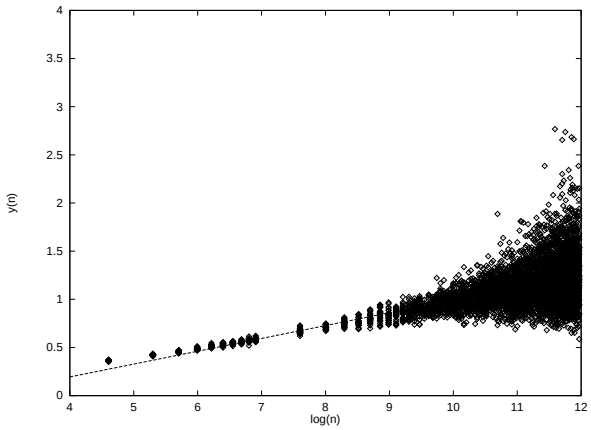

図 2　実験 2 の結果

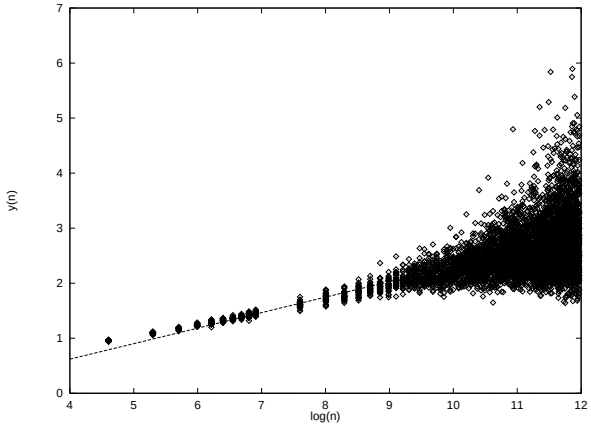

図 3　実験 3 の結果

$$y = 0.1358 \log n - 0.2821.$$

理論値は、$a^* = 0.1666$ である。

(2) 最小二乗法で得られた回帰曲線は

$$y = 0.1325 \log n - 0.3385.$$

分割を変えても大きな変化は見られない。

(3) 最小二乗法で得られた回帰曲線は

$$y = 0.2817 \log n - 0.5071.$$

理論値は、$a^* = 0.3333$ ある。

以上の実験結果を見ると、a^* として、理論値よりもやや小さな値が得られている。

4.3 考　　察

本論で提案した方法は、カルバック情報量 $H(w)$ と事前分布 $\varphi(w)$ で表される学習の順問題において、学習係数が理論または実験で解明されている $H_1(w)$ があるときに、カルバック情報量を

$$H(w) = H_1(w) + H_2(w)$$

と分割する方法を提案した。提案方式で得られた学習係数は、理論値よりも、やや小さめの値であったが、この理由は $H_1(w) = 0$ を満たす w の近傍は $H_2(w) = 0$ を満たす w の近傍よりも広がりを持つためであると考えられる。

本論で提案した方法に関連して、幾つかの観点から考察する。

(1) 確率的複雑さは

$$f(t) = -\log \int \exp(-ntH(w))\varphi(w)dw$$

と置くとき、$f(1)$ に等しい。そこで、

$$\begin{aligned} f(1) &= \int_0^1 \frac{df}{dt} dt \\ &= \int_0^1 dt \Big[\frac{\int nH(w) \exp(-ntH(w))\varphi(w)dw}{\int \exp(-ntH(w))\varphi(w)dw} \Big] \end{aligned}$$

となることを利用して、$f(1)$ を計算する方法がある [6] [7]。ここで、

$$\frac{\int nH(w) \exp(-ntH(w))\varphi(w)dw}{\int \exp(-ntH(w))\varphi(w)dw}$$

は確率分布 $\exp(-ntH(w))$ による $nH(w)$ の平均値であるから、$\exp(-ntH(w))$ からサンプルを取り出して平均することで計算できる。しかしながら、この方法では、t についての積分区間 $[0, 1]$ を原点に近いほど細かく分割し、各 t において確率分布 $\exp(-ntH(w))$ に従うサンプルを取り出す必要があるので、本論文で提案した方法と比較すれば、遥かに演算量が大きい。

(2) 本論で提案した方法が精度よい計算になるためには $\exp(-nH(w))$ と $\exp(-nH_1(w))$ が分布としてできるだけ似ていることが望ましい。そこで、理論的に λ_1 が計算できる $H_1(w)$ の種類を増やしておけば、より多種類のカルバック情報量に対応できることになる。特異点解消定理に基づいてできるだけ多種類の特異点に対応できるようにすることは今後の課題である。

(3) 上記の (1) で述べた方法と本論で述べた方法を組み合わせたハイブリッド方式として

$$g(t) = -\log \int \exp(-ntH_2(w))\rho_n(w)dw$$

と置いて $g(1)$ を次の方法で求めることも考えられる。

$$\begin{aligned} g(1) &= \int_0^1 \frac{dg}{dt} dt \\ &= \int_0^1 dt \Big[\frac{\int nH_2(w) \exp(-ntH_2(w))\rho_n(w)dw}{\int \exp(-ntH_2(w))\rho_n(w)dw} \Big] \end{aligned}$$

この方法は、演算量の点では、本論文で述べた方法よりも大きく、(1) で述べた方法よりは少ない。また a^* を求める精度の観点からは、本論で述べた方法よりも精度は良いものと思われる。

5. 結　　論

カルバック情報量を分割して、学習係数を計算するアルゴリズムを提案し、その有効性を実験的に確認した。今後の課題として、より正確に学習係数が求められるようなカルバック情報量の分割の仕方の最適化の問題がある。

この研究は、科学研究費補助金 15500130 の援助を受けた。

文　　献

[1] 甘利俊一, 尾関智子, 朴慧暎, “階層的モデルにおける学習と推論- 特異構造を持つ統計モデル”, 電子情報通信学会論文誌, Vol.J85-DII, No.5, 701-708, 2002.

[2] 青柳美輝、渡辺澄夫、“特異点解消定理と学習理論への応用,” 信学技報, NC2-26, pp.25-30, 2003.

[3] 伊庭幸人, “マルコフ連鎖モンテカルロ法とその統計学への応用”、統計数理、第 44 巻、第 1 号、49-84,1996.

[4] Yoshihiko Ogata, “A Monte Carlo Method for an objective Bayesian procedure,” Ann. Inst .Statist. Math., Vol.42, No.3, 403-433, 1990.

[5] 永田賢二，渡辺澄夫，“カルバック情報量の分割による特異モデルの学習係数の計算アルゴリズム” 情報論的学習理論 2003 , to appear

[6] 西上功一郎, 渡辺澄夫, “特異な学習モデルの選択における事前分布の影響について” 電子通信情報学会論文誌, J86-D-II, No.1, 119-129, 2003.

[7] 高橋克之, 渡辺澄夫, “特異的な学習モデルにおける MCMC 法の評価,” 情報論的学習理論, 2003.

[8] 渡辺澄夫, “ベイズ法による階層型統計モデルの推定誤差について”, 電子情報通信学会論文誌, Vol.J81-A, No.10, 1442-1452, 1998.

[9] 渡辺澄夫, “特異点を持つ学習モデルと事前分布の代数幾何”, 人工知能学会誌, 16 巻 2 号, 308-315, 2001.

[10] Sumio Watanabe, “Algebraic Analysis for Nonidentifiable Learning Machines”, Neural Computation, 13, 899-933, 2001.

[11] 山崎啓介, 渡辺澄夫, “特異点をもつ推論モデルの学習曲線の確率的計算法”, 電子通信情報学会論文誌 D-II-85J(3), 363-372, 2002.