

## GMDH（発見的自己組織化法）の概要と適用例

——流出問題、水問題を例として——

田 中 雅 史\* 福 間 順\*\* 鳥 井 清 司\*\*

## I. はじめに

現在、環境問題をはじめとしてその解決が望まれているにもかかわらず、対象の持つ複雑さのために近代科学の手法では十分対処しえない問題が数多くある。このような問題の対象は複雑なシステムである場合が多く、その取扱いにはサイバネックスの方法が有効である。ここに複雑なシステムというのは変数やパラメーターが多く、しかもそれらの間に相互の関連性を持ち、しかも多くは非線形な構造を持っているために数学モデルの作成が困難なシステムをさしている。

N. Wiener により創始されたサイバネックスの方法は、『機械および生物体、またはそれらを複合した系における情報の受容、保存、変換および利用の方法』と A. N. Kolmogorov が定義しているように、情報や制御、とりわけ情報と密接な関連を持ち、情報間の変換方式に視点を置く一般的な研究方法であるといえる。したがって、近代科学の分析的性格と対照的に総合的な性格を持ち、問題解決にあたって近代科学の手法では“法則性”を中心として展開されるのに対して、サイバネックスの方法ではもっぱら情報とその変換の過程に注目するという本質的な差異がある。

ところで、われわれが日ごろ対象としている自然のシステムは、生態システムのようにシステム自体に情報過程を持つものではない。しかしながら、開発計画等に見られるように規模が大きくなればなるほど、社会系や生態系をも包含する形で問題解決がせまられている。また、種々の観測系の整備により、より多くの情報が収集されてきている。このような状況のもとで、従来のように情報を単なる法則性を立証する手段としてではなく、より積極的に利用する立場から問題解決を図る方法があってもよいはずである。サイバネックスの効用が、あらゆる分野に共通な一般的方法を提供する点と複雑な問題を解明する点<sup>1)</sup>にあるならば、上述のことからまさに有効な方法であるといえる。しかしながら、サイバネックスの基本的な概念思想はわかっているにもかかわらず具体的な解決法として確立されている方法はそう多くはないし、ま

たあったとしても非常に専門的になっている。

A.G. Ivakhnenko<sup>2), 3)</sup> の提案した発見的自己組織化の原理に基礎を置く Group Method of Data Handling (GMDH) は、このようなサイバネックスの方法を具体的に示したものであり、複雑な問題の解決法として有効である。

GMDH の基本的概念およびアルゴリズムについては池田・榎木<sup>4)</sup>によりきわめて明解で詳細な紹介がなされている。筆者らの2, 3の適用例からみても問題解決の一手法として非常に興味深い点があり、今後、多くの応用性を持つと考え、その概要と適用例についてここに紹介する。

## II. 発見的自己組織化の原理と GMDH

発見的自己組織化 (Heuristic Self-Organization) の原理は人間が対象の認識における知的活動 (問題解決) において通常用いていると考えられる方式である。

すなわち、人間は解決をせまられている種々の問題について自己の経験や学習に基づく何らかの判断基準に従って総合的な見地から対象の持つ有用な情報を選択しつつ試行錯誤的に次第に目的に合致するように解決策を探求していく。このような行為を行ってきた例として動植物の品種改良がある。

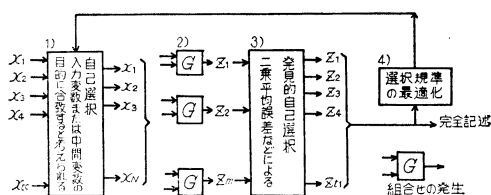
人類は栽培・飼育する動植物の遺伝質を改良して新しい家畜や作物の種類・品種を作り出す行為を有史以来続けている。最初は野生の動物や植物を順化することからはじめ、より優れた品種を得るために、種の進化、生存競争、自然とうたなどを利用して改良を重ねてきた。

この原則を問題解決のための情報処理とモデリングに用いることにより、“複雑さ”を取扱うことが可能となる。これが Ivakhnenko<sup>3)</sup> が提唱している発見的自己組織化法の基礎となっている。

家畜や作物の品種改良では、まず目的とする形質を有する親品種を選び、それぞれを交配して子となる種子を得る。子の品種のうち目的に合わないものはとうたされ、優良なもののみが選択される。子の世代で選択された優良な品種が交配されて孫の世代の種子を得る。このように各世代にわたって、とうたと選択が繰返されて、家畜や作物の新しい品種が完成する。このようにして行

\* 三重大学農学部 (たなか まさふみ)

\*\* 京都大学農学部 (ふくま じゅん, とりい きよし)



- 1) 目的変数（出力）を  $y$  とするとき、 $y$  に関係のありそうなすべての入力変数  $x_1, x_2, \dots, x_{10}$  を考える。この中には系の遅れを考慮した変数も含まれる。この中から互に独立性が強く、かつ出力  $y$  との相関も高い  $N$  個の変数  $x_1, x_2, \dots, x_N$  を選択する ( $N < 10$ )。
- 2)  $N$  個の入力（中間）変数の中から2個ずつ選択して  $NC_2$  個の組合せが発生させ、中間変数  $z_1 = f(x_1, x_2), z_2 = f(x_1, x_3), \dots, z_m = f(x_{N-1}, x_N)$  を作成する。ここに、 $m = NC_2$  である。
- 3) 中間変数  $z_1, z_2, \dots, z_m$  と出力  $y$  との二乗平均誤差の小さい順に  $m$  個（最終的には1個）の中間変数を選択する。
- 4) 最適な中間変数が選択できれば、それを最終的な完全記述式として計算は完了するが、そうでない場合は、選択標準を最適化して1)へ戻る。

図-1 基本的 GMDH の構造

われる品種改良では次のような原則が経験的に確立している。

1. 1つの品種の中に入取れたい形質をそれぞれ持っている原種を選ぶ。原種の選び方によって品種改良を悪化させたり遅らせたりすることもある。
2. 目的とする形質を有する品種を完成させるまでに数世代を要し、複雑な特性を持つ品種を作り出す場合は、それだけ多くのとうたの世代が必要である。
3. 長期にわたる改良は品種の退化をもたらすため、品種改良をやめる最適な世代が存在する。

これらの原則は数学的な厳密さの意味では証明されていないので“品種改良の仮説”とされている<sup>4)</sup>。この仮説を利用して、発見的自己組織化のアルゴリズムを作成することは可能である。

育種学の分野においては、品種改良の方法として、非常に多くのものが体系づけられているが<sup>5)</sup>、GMDHにおいても同じように多くの方法が考えられる。そのうち基本的な GMDH のアルゴリズムの構造を図示したのが図-1 である。

このようなアルゴリズムを簡単な関数の例で示してみよう。まず出力  $y$  とそれに関係のありそうな4つの入力変数  $x_1, x_2, x_3, x_4$  を想定する。ここで、 $x_1 = 1, x_2 = p, x_3 = p^2, x_4 = p^3, y = 1 + p + p^2 + p^3$  のように与え、 $p$  の値が  $p = -9 \sim 10$  の間で1おきに变化すれば、各変数の系列は表-1 に示したような値をとる。

このような観測値が得られた時、出力  $y$  が GMDH によって、どのように再現されていくかを追跡してみる。

数列をトレーニングデータとチェックングデータの2つに分割する。ここでは奇数番目を前者に、偶数番目を

表-1 簡単な関数の例

No.	$X_1$	$X_2$	$X_3$	$X_4$	$Y$
トレーニングデータ	1	1	-9	81	-729
	3	1	-7	49	-343
	5	1	-5	25	-125
	7	1	-3	9	-27
	9	1	-1	1	-1
	11	1	1	1	1
	13	1	3	9	27
	15	1	5	25	125
	17	1	7	49	343
チェックングデータ	19	1	9	81	729
	2	1	-8	64	-512
	4	1	-6	36	-216
	6	1	-4	16	-64
	8	1	-2	4	-8
	10	1	0	0	0
	12	1	2	4	8
	14	1	4	16	64
	16	1	6	36	216
	18	1	8	64	512
	20	1	10	100	1000

後者に割当てている。中間変数を発生させるために、次のような線形の部分記述式を考える。

$$f(x_i, x_j) = a_0 + a_1 \cdot x_i + a_2 \cdot x_j$$

4個の入力変数からは6個の中間変数を発生させることができる。これらすべてについて、その係数  $a_0, a_1, a_2$  を、トレーニングデータを用いて最小二乗法的に求め、チェックングデータを用いてその精度（二乗平均誤差）を計算すれば次のようになった。

$$z_4 = 1 + 59.60 x_2 + x_3 \quad \epsilon_4^2 = 171.4$$

$$z_6 = 1 + x_3 + 1.015 x_4 \quad \epsilon_6^2 = 2.519$$

第二層目では、新たに  $z_4, z_6$  を入力変数として、次のような結果を得た。

$$y = -0.08291 + 0.01676 z_4 + 0.9857 z_6 \quad \epsilon^2 = 0.0742$$

すなわち、第二層目で最適な複雑さを持つ同定式に達し、誤差のレベルも 1/30 以下に減少した。

第一層目の  $z_4, z_6$  を第二層目の部分記述式に代入して完全記述式を作れば、

$$y = 0.91955 + 0.99889 x_2 + 1.00246 x_3 + 1.00048 x_4$$

となる。これから、ほとんど正しくもとの多項式を再現していることが理解できる。

以上のように、GMDH は多くの入力変数の中から、関係のある変数を選び出す機能と、その変数から構成される多項式の中で、最適な複雑さを持つ構造を作り出す機能を備えているといえる。

### III. GMDH のアルゴリズム

多変数入力  $x_1, x_2, x_3, \dots, x_N$  と、それに対応する出力  $y$  からなる1つのシステムがある。このシステムにつ

表-2 入力変数  $x_j$  と出力  $y$  の図式

$i \backslash j$	$x_1$	$x_2$	$x_3$	.....	$x_j$	.....	$x_N$	$y$
$x_1$	$x_{11}$	$x_{12}$	$x_{13}$	.....	$x_{1j}$	.....	$x_{1N}$	$y_1$
$x_2$	$x_{21}$	$x_{22}$	$x_{23}$	.....	$x_{2j}$	.....	$x_{2N}$	$y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$x_{i1}$	$x_{i2}$	$x_{i3}$	.....	$x_{ij}$	.....	$x_{iN}$	$y_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_k$	$x_{k1}$	$x_{k2}$	$x_{k3}$	.....	$x_{ki}$	.....	$x_{kN}$	$y_k$

いて表-2 に示すように  $K$  個の離散的な時刻において、観測値が得られている時、これらのデータを用いて、任意の時刻  $t$  ( $1 \leq t \leq k$ ) における出力  $y(t)$  を推定しようとする問題を内挿問題 (Interpolation Problem) という<sup>6)</sup>。

この内挿問題を解くためには、

$$y(t) = F(x_1, x_2, x_3, \dots, x_j, \dots, x_N) \quad (1)$$

で表現される決定関数 (Decision Function) を求めることが必要である。このように 1 つの系に対応した決定関数を決めてやることを系の同定 (Identification) という<sup>7)</sup>。

今、表-2 に示すようなシステムを、いわゆる Black Box 的認識によって評価するとすれば、システムを非線形として、核関数 (= 応答関数)  $h_n$  を使って、Volterra 型の級数展開<sup>8)</sup> により、その一般的な形は次のように書ける。

$$y(t) = \sum_{n=0}^N \int \dots \int_0^\infty h_n(\tau_1, \tau_2, \dots, \tau_n) \prod_{j=1}^n x(t-\tau_j) \cdot d\tau_1 \cdot d\tau_2 \cdot \dots \cdot d\tau_n \quad (2)$$

ここで、 $N=1$  とし、 $h_0=0$  とすれば

$$y(t) = \int_0^\infty h_1(\tau_1) x(t-\tau_1) d\tau_1 \quad (3)$$

となって、通常われわれが用いる 1 階のたたみこみ積分 (Convolution Integral) である。

いま、 $N=2$  とすれば 2 階までの、 $N=3$  とすれば 3 階までのたたみこみ積分の総和によって、システムを表現していると考えることができる。したがって、 $y(t)$  なるシステムは線形システムと順次高次の非線形性を持つシステムが並列的に結合されたものとして理解でき<sup>9)</sup>、これを視覚的に図示したのが 図-2 である。ここで、 $f_1(t)$  は線形のたたみこみ積分を表わし、 $f_2(t)$  は 2 階の、 $f_i(t)$  は  $i$  階のたたみこみ積分を表現している。

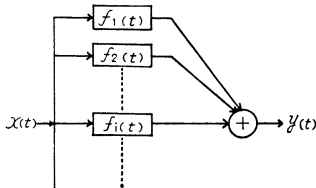


図-2 システムのブロックダイアグラム

このことから、より高階のシステムの全出力  $y(t)$  に対する寄与が小さくなることがわかっておれば、非線形システムの出力は有限な階までのたたみこみ積分の組合せにより、所望の精度で近似できる。

連続関数のかわりに、入出力値の離散値をとると(2)の離散形として、次式を得る。

$$y(t) = \sum_{\tau_1=1}^M h_1(\tau_1) x(t-\tau_1) + \sum_{\tau_1=1}^M \sum_{\tau_2=1}^M h_2(\tau_1, \tau_2) x(t-\tau_1) \cdot$$

$$x(t-\tau_2) + \sum_{\tau_1=1}^M \sum_{\tau_2=1}^M \sum_{\tau_3=1}^M h_3(\tau_1, \tau_2, \tau_3) \cdot x(t-\tau_1) \cdot$$

$$x(t-\tau_2) x(t-\tau_3) + \dots, \dots \quad (4)$$

上式において  $x(t-\tau_1) = x_1$ ,  $x(t-\tau_2) = x_2$ , ... とし、 $h_1(\tau_1) = a_i$ ,  $h_2(\tau_1, \tau_2) = a_{ij}$ , ... とおけば、

$$y(t) = \sum_i a_i x_i + \sum_{i,j} a_{ij} x_i x_j + \sum_{i,j,k} a_{ijk} x_i x_j x_k + \dots, \dots \quad (5)$$

これは Kolmogorov - Gabor の多項式である。すなわち、(1)式の間数  $F$  は所望の精度で(5)の多項式により近似できる。

ここで次のような簡単な形の関数を考えると、

$$f(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \quad (6)$$

(1)式が高次の多項式であるならば、次のようないくつかの部分記述式によって置きかえられるはずである。

$$z_1 = f(x_1, x_2), z_2 = f(x_1, x_3), \dots, z_m = f(x_{N-1}, x_N) \quad (7.1)$$

ここに、 $m = {}_N C_2$

$$v_1 = f(z_1, z_2), v_2 = f(z_1, z_3), \dots, v_p = f(z_{m-1}, z_m) \quad (7.2)$$

ここに、 $p = {}_m C_2$

$$w_1 = f(v_1, v_2), w_2 = f(v_1, v_3), \dots, w_q = f(v_{p-1}, v_p) \quad (7.3)$$

ここに、 $q = {}_p C_2$

(1)式が線形多項式であるならば(6)式の代わりに次のような関数を用いればよい。

$$f(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j \quad (8)$$

また、(1)式が非線形であるが高次の項を含まない場合には、

$$f(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j \quad (9)$$

のような関数形を用いればよい。この(6), (8), (9)式のような関数を基礎関数 (reference function) という。

入力変数  $x_1, x_2, x_3, \dots, x_N$  から中間変数  $z_1, z_2, z_3, \dots, z_m$ , 続いて  $v_1, v_2, v_3, \dots, v_p$ , 続いて……と変換していくにしたがって、近似度は高くなり、式の複雑さも増していく。この際の変換係数の求め方は以下の注)に示す。

注) ガウスの正規方程式系について

二つの独立変数  $x$  および  $y$  について、有限個の値  $x = x_i, y = y_i$ , ( $i = 1, 2, \dots, K$ ) に対応する関数値  $z_i = f(x_i, y_i)$  が与えられている時、 $f(x_i, y_i)$  が次のような二次の多項式で表現できるとすれば、

$$f(x_i, y_i) = a_0 + a_1 x_i + a_2 y_i + a_3 x_i y_i + a_4 x_i^2 + a_5 y_i^2$$

係数  $a_0, a_1, \dots, a_5$  は次のような手順で求めることができ、かつその値の精度をチェックすることができる。

$x_i, y_i, z_i$  ( $i = 1, 2, \dots, K$ ) の既知な  $K$  個の系列を  $1, \dots, K_t$  と  $K_t + 1, \dots, K$  の二つに分割する。(ただし  $6 < K_t < K$ )

前者の系列を用いて、残差  $e_i = z_i - f(x_i, y_i)$  の平方和

$$\|e\|^2 = \sum_{i=1}^{K_t} e_i^2$$

が最小となるように、すなわち、最小二乗近似的に  $a_0, a_1, \dots, a_5$  を求める。すなわち、 $\|e\|^2$  を  $a_i$  ( $i = 0, \dots, 5$ ) について偏微分し、その結果を 0 において極値を求める方程式系を作成し、それを解けばよい。

$\|e\|^2$  をすべて書けば、

$$\|e\|^2 = [z_1 - f(x_1, y_1)]^2 + [z_2 - f(x_2, y_2)]^2 + \dots + [z_{K_t} - f(x_{K_t}, y_{K_t})]^2$$

となる。これから、 $\partial\|e\|^2/\partial a_0 = 0, \partial\|e\|^2/\partial a_1 = 0, \dots, \partial\|e\|^2/\partial a_5 = 0$  を導けば、

$$\begin{pmatrix} K_t & \sum_{i=1}^{K_t} x_i & \sum_{i=1}^{K_t} y_i & \sum_{i=1}^{K_t} x_i y_i & \sum_{i=1}^{K_t} x_i^2 & \sum_{i=1}^{K_t} y_i^2 \\ \sum_{i=1}^{K_t} x_i & \sum_{i=1}^{K_t} x_i^2 & \sum_{i=1}^{K_t} x_i y_i & \sum_{i=1}^{K_t} x_i^2 y_i & \sum_{i=1}^{K_t} x_i^3 & \sum_{i=1}^{K_t} x_i^2 y_i \\ \sum_{i=1}^{K_t} y_i & \sum_{i=1}^{K_t} x_i y_i & \sum_{i=1}^{K_t} y_i^2 & \sum_{i=1}^{K_t} x_i y_i^2 & \sum_{i=1}^{K_t} x_i^2 y_i & \sum_{i=1}^{K_t} y_i^3 \\ \sum_{i=1}^{K_t} x_i y_i & \sum_{i=1}^{K_t} x_i^2 y_i & \sum_{i=1}^{K_t} x_i y_i^2 & \sum_{i=1}^{K_t} x_i^3 y_i & \sum_{i=1}^{K_t} x_i^2 y_i^2 & \sum_{i=1}^{K_t} x_i y_i^3 \\ \sum_{i=1}^{K_t} x_i^2 & \sum_{i=1}^{K_t} x_i^3 & \sum_{i=1}^{K_t} x_i^2 y_i & \sum_{i=1}^{K_t} x_i^4 & \sum_{i=1}^{K_t} x_i^3 y_i & \sum_{i=1}^{K_t} x_i^2 y_i^2 \\ \sum_{i=1}^{K_t} y_i^2 & \sum_{i=1}^{K_t} x_i y_i^2 & \sum_{i=1}^{K_t} y_i^3 & \sum_{i=1}^{K_t} x_i^2 y_i^2 & \sum_{i=1}^{K_t} x_i y_i^3 & \sum_{i=1}^{K_t} y_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{K_t} z_i \\ \sum_{i=1}^{K_t} z_i x_i \\ \sum_{i=1}^{K_t} z_i y_i \\ \sum_{i=1}^{K_t} z_i x_i y_i \\ \sum_{i=1}^{K_t} z_i x_i^2 \\ \sum_{i=1}^{K_t} z_i y_i^2 \end{pmatrix}$$

となる。これがガウスの正規方程式系である。この方程式を解いて、 $a_0, a_1, \dots, a_5$  を求めれば、関数  $f(x_i, y_i)$  が定まる。ここで、 $a_0, a_1, \dots, a_5$  を求めるために用いた独立変数の系列  $x_i, y_i$  ( $i = 1, \dots, K_t$ ) の組を、その果たす役割に従って、トレーニングデータといい、トレーニングデータの統計的な特性を係数  $a_i$  ( $i = 0, \dots, 5$ ) に記憶させているとも考えられるから、ここまでの過程を学習過程ともいう。

求められた関数の精度を別のデータ系列  $x_i, y_i$  ( $i = K_t + 1, \dots, K$ ) を用いて二乗平均誤差の値で検討する。

$$\|e\|_{\text{check}}^2 = \frac{1}{K - K_t} \sum_{i=K_t+1}^K [z_i - f(x_i, y_i)]^2$$

ここで用いられるデータを上述のトレーニングデータに対して、チェックデータといい、この過程をテスト過程という。テスト過程において誤差の小さい関数を正則性 (regularity) が高いという。正則性が低い関数、すなわち  $\|e\|_{\text{check}}$  の大きい関数は有害な中間変数を発生させる場合が多い。

最終的に得られる最適な近似度を持つ多項式は次のように定義する。

(7.1), (7.2), (7.3),  $\dots$ , の各層で発生させた中間変数  $z_j, v_j, w_j, \dots$  と出力  $y$  との二乗平均誤差の各層における最小値をつぎのように表現する。

$$E_1^2 = \min. \frac{1}{K} \sum_i (y_i - z_{ij})^2 \quad j = 1, 2, \dots, m \quad (10.1)$$

$$E_2^2 = \min. \frac{1}{K} \sum_i (y_i - v_{ij})^2 \quad j = 1, 2, \dots, p \quad (10.2)$$

$$E_3^2 = \min. \frac{1}{K} \sum_i (y_i - w_{ij})^2 \quad j = 1, 2, \dots, q \quad (10.3)$$

$E_1, E_2, E_3, \dots$ , は一般的には次第に減少していき、ある層で最小値に達し、その層以後は急激に増大する。

$$E_1 > E_2 > E_3 > \dots > E_k < E_{k+1} \quad (11)$$

となることが経験的に知られている。この  $k$  層において最小の二乗平均誤差を持つ中間変数をもって、最適な複雑さを持つ出力  $y$  の完全記述式という。

この近似式がシステムの同定式としては最善であるが、しかし、システム出力の予測式としても十分安定した正しい予測を行うかということになれば、必ずしもそうではない。正しい予測を行うためには、正則性 (regularity) が高いことが必要である<sup>7)</sup>。ここでいう正則性とは、標本で作成した同定式が母集団全体についても成立つことを意味している。

正則性が高く、かつできるだけむだな計算を省いた方法で完全記述式を得るために、次の2つの手順を付け加える。

最初、表-2に掲げた観測値を  $(X_1, y_1), (X_3, y_3), \dots, (X_{2i+1}, y_{2i+1})$  という奇数番目のグループと  $(X_2, y_2), (X_4, y_4), \dots, (X_{2i}, y_{2i})$  という偶数番目のグループの2つに分け、一方を(7)の中間変数  $z_j, v_j, w_j, \dots$  を発生させる係数  $a_0, a_1, a_2, \dots, a_5$  を求めるために用い、他方を求めた係数の精度を調べるために用いる。前者をトレーニングデータといい、後者をチェックデータという。一般的には分散値の高い方をトレーニングデータに用い、低い方をチェックデータに用いている。

いま一つは不必要な計算を省くことであり、(7)式で示しているように、 $N$  個の入力変数があれば、層を通過するにしたがって、中間変数の数は、

$$m (= {}_N C_2) < p (= {}_m C_2) < q (= {}_p C_2) < \dots$$

と飛躍的に増大していき、膨大な計算が必要になる。そこで第一層目については  ${}_N C_2$  個の組合せすべてについて計算するが、第二層目以後は、二乗平均誤差の小さい順に  $l$  個選んでやれば、二層目からは常に  ${}_l C_2$  個の組合せについてのみ計算すればよく、大幅に計算量を少なくすることができる。

この  $l$  の値については、大きくとりすぎると計算の能率を落とす結果になり、小さくとりすぎると近似の精度を犠牲にすることになるため、注意深い配慮が必要である。普通は入力変数の数と同程度の大きさととればよい。

これらの計算の流れを示したのが 図-3 のフローチャートである。流れに沿って説明すれば、

- ① 観測データの読み込み
- ② 読み込んだデータを奇数番目のグループと偶数番目のグループの2つに分け、分散値の大きい方をトレーニングデータ、小さい方をチェックデータとする。
- ③ 通過層数のカウント
- ④, ⑤ 組合せ ( $i, j$ ) の発生
- ⑥ 組合せ数のカウント
- ⑦ 2つの入力変数  $x_i, x_j$  のトレーニングデータを用い

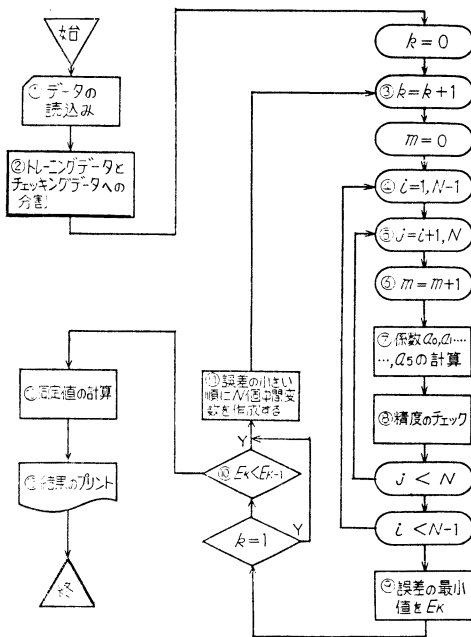


図-3 GMDH アルゴリズムのフローチャート

て、最小二乗法的に変換係数  $a_0, a_1, \dots, a_5$  を求める。

- ⑧  $x_i, x_j$  のチェックングデータを用いて、中間変数と出力との二乗平均誤差  $\varepsilon_m$  を計算する。
- ⑨  $\varepsilon_m, m=1, \dots, M$  の値の小さい順に並べかえ、その最小値を  $E_k$  とする。
- ⑩  $k \geq 2$  ならば  $E_{k-1}$  と比較し  $E_k < E_{k-1}$  ならば次の層へ進む。
- ⑪  $\varepsilon_m$  の小さい順に  $N$  個対応する中間変数  $z_m$  を選び、 $x_i$  におきかえる。
- ⑫  $E_k \geq E_{k-1}$  ならば計算は終了する。
- ⑬ 同定値、各層において中間変数を作り出した係数、その誤差等の必要な結果を出力する。

#### IV. 適用例

GMDH は、関係があると思われるがその因果関係が明らかでない多くの変数の中から、発見的に重要な変数を捜して、最適な複雑さを持つ多項式を作り出す機能を持っている。したがって、本質的に複雑なシステム、たとえば、人間社会の経済活動、環境問題、生態系の諸問題など、たんに変数が多いというだけでなく、それらが動的にからみ合った複雑なシステムが持つさまざまな問題を解決するために適用することができる<sup>10)</sup>。

A. G. Ivakhnenko が GMDH を適用した具体的な例としては、「小麦の作付面積の予測<sup>11)</sup>」, 「貯水池に生息す

るバクテリア量の予測<sup>12)</sup>」, 「イギリスの経済予測<sup>13)</sup>」などがある。

わが国では、池田<sup>4)</sup>らによって詳しく紹介されて以来、徐々に研究されつつある。市川<sup>13)</sup>は河川の水質問題に適用しており、池田<sup>14)</sup>らはアルゴリズムを修正して流出量予測を行っている。田中<sup>15)</sup>は貯水池からの放流水中の濁度の予測にこの方法を応用している。

ここでは、以下で木津川上流域の流出問題と木曽川の濁度の問題に対して適用した例について述べる。

##### 1. 流出問題の例

流域内の数カ所で観測された降雨量と流域下流端から流出する流量が測定されている時、流出量が GMDH アルゴリズムの適用によってどのような構造で同定されるかを検討した。データとして三重県伊賀上野地方の木津川流域内(流域面積 512km<sup>2</sup>)で観測された 5 地点の降雨量  $A, B, C, D, E$  (単位 mm/day) と、流域末端で測定された流量  $Q$  (m<sup>3</sup>/sec) が、昭和 44 年 5 月 1 日～6 月 30 日までの 2 カ月間にわたって与えられている。

GMDH アルゴリズムの適用の手続きは次のようである。

- i) 入力変数としては  $A, B, C, D, E$  の 5 地点における当日から 5 日前までの降雨量と流域末端での前日から 5 日前までの流量  
( $A_0, A_1, \dots, A_5, B_0, B_1, \dots, B_5, C_0, C_1, \dots, C_5, D_0, D_1, \dots, D_5, E_0, E_1, \dots, E_5, Q_1, Q_2, \dots, Q_5$ ) の 35 変数とした。
- ii) トレーニングデータとチェックングデータの分割は奇数番目を前者へ、偶数番目を後者へ割当てた。
- iii) 中間変数は流出量との二乗平均誤差の小さいものから順に 35 個選択した。
- iv) 中間変数を発生させる式としては(6)式を用いた。

以上の手続きの結果得られた構造を図-4 に示す。

すなわち、第二層目で最適な複雑さに達し、 $k$  日目の流出量  $Q_k$  は、 $k-1$  日目の流量  $Q_{k-1}$  と  $B$  地点での降雨量  $B_{k-1}$ 、 $k-3$  日目の  $E$  地点での降雨量  $E_{k-3}$ 、 $k-5$  日目の  $C$  地点での降雨量  $C_{k-5}$  から構成されるものが最適値を示した。その時の部分記述式および完全記述式は次のようになる。

Layer 1 (部分記述式) :

$$y_1 = 3.032 + 0.7605 Q_1 - 0.8307 C_5 + 0.08606 Q_1 C_5 \\ + 0.001788 Q_1^2 + 0.006288 C_5^2 \\ y_2 = 6.215 + 2.212 B_1 - 0.07925 E_3 + 0.001544 B_1 E_3 \\ - 0.02206 B_1^2 + 0.006564 y_2^2$$

Layer 2 (完全記述式) :

$$Q_0 = 3.336 + 0.355 y_1 - 0.3496 y_2$$

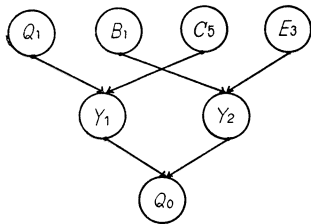


図-4 最適な同定式の構造

$$+0.04314y_1y_2-0.004167y_1^2 \\ +0.006564y_2^2$$

この式による流出量の同定値と実際の観測値との比較を示したのが、図-5である。この時の両者の相関係数は0.958、平均二乗誤差は8.245であり、非常によい一致を示した。

## 2. 水質問題の例

出水時における木曽川の濁度の時間的変化について検討した。解析に用いたデータは、昭和50年8月23日8時から24日18時に至る期間の測定記録を2時間ごと（単位時間）に読取り用いた。流量データは犬山地点における $Q-H$ 曲線を用いて水位から算出した。

入力変数としては以下のように4単位時間前までの流量と単位時間前までの濁度を考慮した

$$T_i = f(Q_{i-4}, Q_{i-3}, Q_{i-2}, Q_{i-1}, Q_i, T_{i-3}, T_{i-2}, T_{i-1})$$

ここに、 $T_i$ ： $i$ 単位時間の濁度、 $Q_{i-k}$ ： $i$ 単位時間より $k$ 単位時間前の流量、 $T_{i-k}$ ： $i$ 単位時間より $k$ 単位時間前の濁度である。

図-6に決定された3つのモデルによる同定結果を示した。図中の○印は実測値を示す。

**モデル-1** 計算上の最適モデルで、この構造式は、

$$T_i = f(Q_{i-3}, Q_{i-2}, Q_{i-1}, T_{i-3}, T_{i-2}, T_{i-1})$$

なる6変数からなるきわめて複雑な形をしており、最高べき乗は32乗にもなる。注目されるべきことは、 $i$ 単位時間の濁度はそれより前の時間の変数で決定されていることで、以下のモデルにおいてもすべて共通していえる。すなわち、換言すれば1単位時間先の濁度が予測できることを示唆している。この場合の相関係数は0.997であった。

**モデル-2** 基礎関数として線形の(8)式を採用したもので、てい減部を除いて、よく合っている。その構造式は、

$$T_i = 0.006 + 0.13276 Q_{i-1} + 0.63929 T_{i-1}$$

である。ここに、 $Q_{i-1}$ は実際の流量を1/1000倍、 $T_i$ 、 $T_{i-1}$ については1/100倍してある。これは数値のオー

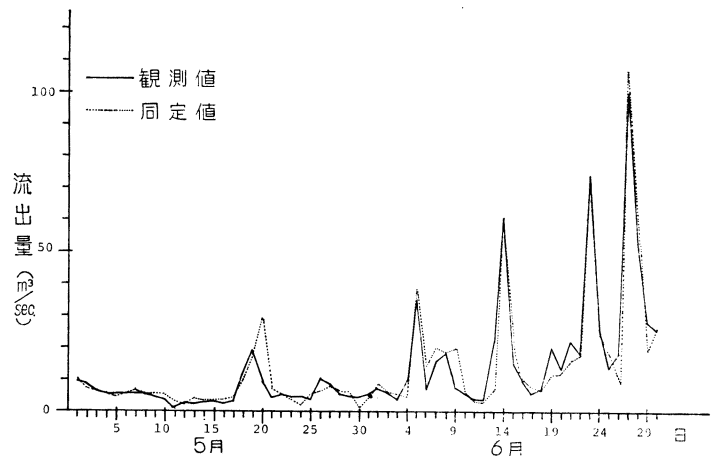


図-5 観測値と同定値の比較（流出問題の例）

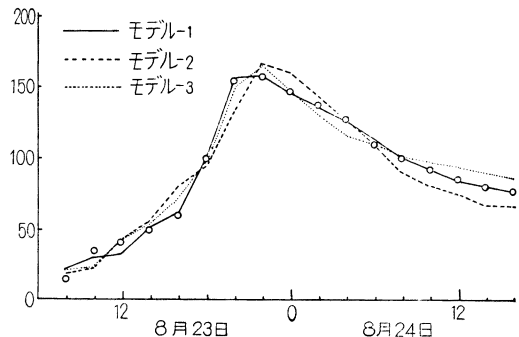


図-6 観測値と同定値の比較（水質問題の例）

ダーをそろえて、係数値を見やすくしたためで、結果に実質的な影響は及ぼさない。以下のモデルについても同様の配慮をしてある。上式から単位時間内の濁度変化量 $\Delta T_i$ は、

$$\Delta T_i = T_i - T_{i-1} = 0.006 + 0.13276 Q_{i-1} - 0.36071 T_i$$

となる。この場合の相関係数は0.966であった。

**モデル-3** 基礎関数として非線形の(6)式を採用したもので、その構造式は、

$$T_i = 0.2038 - 0.19231 Q_{i-1} + 1.2051 T_{i-1} \\ - 0.064925 Q_{i-1} \cdot T_{i-1} + 0.061296 Q_{i-1}^2 \\ - 0.21034 T_{i-1}^2$$

であり、相関係数は0.984であった。

**モデル-2**に比して構造が複雑になっただけ精度はあがるが、てい減部では合わない。したがって、**モデル-1**のように $i$ 時間における濁度はそれ以前の流量、濁度に支配されているのが実態であろう。**モデル-1**の結果は、その時間スケールが3単位時間（6時間）位であることを示している。

## V. ま と め

以上2つの例題について GMDH の応用例を示してきたが、これらを通じていえる GMDH の特徴は次のようである。

1) 複雑な機構を持つ系の数学モデルを見出そうとした時、いきなり完全記述式を解くのではなく、部分記述式を解いて中間変数を発生させるという段階を重ねて完全記述式に至っているため、部分記述式を解くのに必要なだけの短いデータがあれば適用可能である。

2) データの観測手段は完全なものではなく、それゆえ入力変数や中間変数は必ずしも有用な情報ばかりであるとは限らず、時には有害なものも含むこともありうる。しかし、これらの有害な情報は中間変数を選択する過程で自動的に排除され有益な情報のみが次の層へ伝わる。それゆえ、非常によい精度の同定結果が期待できる。

3) たとえ非常に長いデータが利用できたとしても、完全記述式に基づいた正規方程式を解くのではないために大容量の計算機を用いずとも、小型のものでも十分計算可能である。

4) 正規方程式の係数行列は常により条件の行列とはいえない。しかし、多くの部分記述式の中から常により条件を具備した行列を見出すことは可能であり、これはまた最終的に得られる完全記述式の正則性を保つためにも役立つ。

多変数で、その物理的構造が明確でなく、かつそれに関する情報量が少ない複雑な系の数学モデルを見出そうとした時、GMDH は非常に有効な手段を提供する。

しかも、そのモデルの作成過程において、正則性を保つような考慮が払われているために、将来の入力データについても、かなりの精度の予測が可能である。

今回は流出問題と水質問題のみに適用を試みたが、この方法は柔軟性に富み、これら以外の分野への応用も十分に考えられて、今後の発展が大いに期待できる手法であると思われる。

### 引用文献

- 1) W. Ashby: サイバネティクス入門 (篠崎ほか訳), 宇野書店 (昭和42年)
- 2) A. G. Ivakhnenko: The Group Method of Data Handling—A Rival of the Method of Stochastic Approximation, Soviet Automatic Control, 13 (3), pp. 43~55, (1968)
- 3) A. G. Ivakhnenko: Heuristic Self-Organization in Problems of Engineering Cybernetics, Automatica Vol. 6 pp. 207~219 (1970)
- 4) 池田・樺木: GMDH (発見的自己組織化法) と複雑な系の同定・予測, 計測と制御, 14 (2), pp. 185~195 (1975)
- 5) たとえば, 体系農業百科事典第II巻, 東畑精一ほか監修, 編集発行財団法人農政調査委員会, pp. 79~122 (1966)
- 6) 古田勝久: 線形システムの観測と同定, コロナ社, p. 98 (1976. 1)
- 7) A. G. Ivakhnenko et al: Regularization of Decision Functions in the Group Method of Data Handling, Soviet Automatic Control, 15 (2), (1970)
- 8) Volterra, V: Theory of Functionals of Integral and Integro-Differential Equations Dover (1959)
- 9) J. Amorochio et al: Determination of Nonlinear Functional Response Functions in Rainfall-runoff Processes, Water Resources Research, 7 (5) pp. 1087~1101
- 10) 合田周平: 予測の科学, 講談社ブルーバックス, pp. 96~98 (1976. 11)
- 11) A. G. Ivakhnenko et al: Mathematical Simulation of Complex Ecological Systems, Soviet Automatic Control 4 (4) pp. 15~26 (1971)
- 12) A. G. Ivakhnenko: Structural Identification of Differential Equations by a Small Number of Observations Using Self-Organization Method, Soviet Automatic Control 5 (2) pp. 31~44 (1972)
- 13) 市川新: 発見的自己組織化法による水質予測, 第20回水理講演会論文集, (1976. 2)
- 14) 池田・落合・樺木: 逐次GMDHアルゴリズムとその河川流量予測への応用, 計測自動制御学会論文集, 12 (2) (1976)
- 15) 田中雅史: 人造貯水池からの放流水中の濁度推定法について, 三重大学農学部学術報告, 第50号, pp. 25~30 (1975. 12)

[1976. 11. 10. 受稿]

(p. 76 より)

八郎湯干拓事業概要 (英訳付) (東北農政局八郎湯新農村建設事業団)

林地の水および土壌保全機能に関する研究 (第2報)

地床かく乱が地表流下・浸透および侵食に及ぼす影響と林地の保全策 (村井 広・岩崎勇作)

石川県の水理地質と地下水 (北陸農政局計画部)

福井県の水理地質と地下水 ( )

土の物理性と土壌診断 (三好 洋・丹原一寛)

第14回理工学における同位元素研究発表要旨集 (第14回理工学における同位元素研究発表会運営委員会)

新潟平野の地盤沈下 本文 (北陸農政局信濃川水系農業水利調

査事務所)

新潟平野の地盤沈下 添付資料Ⅰ 一般解析編 ( )

新潟平野の地盤沈下 添付資料Ⅱ 時系列 (トレンド) 解析編 ( )

新潟平野の地盤沈下 添付資料Ⅲ 地下水流動 (圧密) 解析編 ( )

計画基準作成調査報告書 (公害対策調査) 西川地区 (中国四国農政局計画部)

Scientia Sinica: Vol. 20, No. 1; 1977 (中国)

国立台湾大学農学院研究報告: 17巻1号; 1977 (台湾)

建設技術: 4巻3号, 5号, 6号; 1977 (韓国)

(p. 36へ)

上田の重粘土水田における暗キヨ排水が土壌および水稻生育に及ぼす影響

国分欣一・増島 博・根本清一・長野間宏・清井敏博

洪積層に由来する長野県上田市の重粘土水田において、各種暗キヨの排水効果と水稻生育収量への影響について検討し、暗キヨ施工によって中干しの効果が発揮されることを認めた。その内容は時期別排水量、中干し時の地下水位、落水後の土壌三相、貫入抵抗の変化などである。

(農土誌 45-8, pp. 5~10, 1977)

フレキシブルパイプ使用による water hammer の軽減対策

弥永 孝一・西山 壮一・田地野直哉

農業用管水路の配管、管内圧力、流速等は他の目的の管路におけるそれらとは通常異なる。したがって、農業用管水路に適した独自の water hammer の軽減方法を考えねばならない場合がある。そこで筆者らは、より簡単な軽減方法としてパイプラインの一部にフレキシブルパイプを使用した water hammer の軽減対策を考えた。それについて室内実験を行い、考察を試みた。

(農土誌 45-8, pp. 11~14, 1977)

GMDH (発見的自己組織化法) の概要と適用例

田中 雅史・福岡 順・鳥井 清司

環境問題等の、その物理的構造が明確でなく、かつそれに関する情報量が少ない系に関する問題の取扱いに関しては、GMDH は非常に有効な方法である。GMDH は、A. G. Ivakhnenko によって提案されたもので、サイバネテックスの方法を具体的に示したものである。

ここでは、その基礎的な概念とそのアルゴリズムを説明し、その適用例を流出問題と水質問題について示した。

(農土誌 45-8, pp. 15~21, 1977)

暗キヨ埋戻し土塊の安定性

神部 広之・光永 演允・堀 要・杉山 英雄

粘質土壌地帯の暗キヨ排水においては、暗キヨ埋戻し部の役割が重要だといわれている。これは、埋戻し部の土塊の間を水が通過するため、この土塊を高分子系土壌改良剤 (暗キヨ排水機能促進剤) によって保護し、暗キヨ機能を持続させるのが有効であることは既に指摘した。ここでは現場適用の基礎試験として、土性、土塊の大きさ、含水比によって土塊の保護効果の差異を検討し、さらに土塊の崩壊要因についても検討し報告している。

(農土誌 45-8, pp. 22~26, 1977)

×

×

×

×