対話形 GMDH による広域パターンの同定[†]

田村 坦之*・青谷 年永*

Large-Spatial Pattern Identification by Conversational GMDH

Hiroyuki TAMURA* and Toshinaga AOTANI*

As a new application of GMDH, a large-spatial pattern identification problem is solved where space is the independent variable of the system. The interest to solve this problem has been stimulated by the practical situation of monitoring air pollution concentration. Selection of input variables is one of the most important tasks in the GMDH. Two kinds of input variables are tested for our purpose:

Type 1: (x, y) coordinates

Type 2: (x, y) coordinates and the output of pointwise information at the neighboring points of a prediction point.

Although Type 2 involves more computational tasks than Type 1, we can obtain better accuracy of pattern identification from Type 2.

GMDH algorithm is implemented in the Time Sharing System (TSS), therefore we could use the computer program by the man-computer interactive way which enables us to change some heuristics in the GMDH easily.

1. ま え が き

システムに関する情報が不足していたり、入力変数と出力変数のつながりが複雑で完全な数式モデルを作ることができない場合、既知の入出力データのみからシステムをモデリングする方法の一つに自己組織化法に基づいたGMDH(Group Method of Data Handling)がある^{1)~3)}。GMDHを予測モデルの構成のために用いると、従来の重回帰手法による線形回帰モデルと同形のもののみならず、非線形モデルが容易に作れる。また、モデリングに必要なデータの数はいくらでも少なく済む利点がある。GMDHは、これまで主に、独立変数に時間を取り、システムの未来値を予測する問題に適用されてきた^{4), 5)}。

本論文では、時間を固定し、独立変数に空間座標をとり空間的に広がりをもつ広域系のパターン同定にGMDHを適用する方法を検討する。さらに、GMDHを適用して自己組織化を行う場合に必要となる種々のヒューリスティクスおよびしきい値の変更を迅速にかつ容易に行うために、大型計算機のTSS機能を利用した対話形式のGMDHを用いる。

以下、2章でGMDHの発見的自己組織化(Heuristic Self-Organization)の原理と多層構造のアルゴリズムを概説し、3章ではGMDHを応用して広域系のパターンを同定する方法を示す。Ivakhnenkoらのやり方⁶⁾を使ったType 1の方法に対し、新たにType 2の方法を提案する。4章では、数値例として2種類の広域パターンを想定し、3章で提案した方法を適用してその推定精度の比較を行う。

2. GMDH の原理とアルゴリズム

GMDHの一般原理と応用については、最近、すぐれた解説³⁾が出されているが、ここでは、本論文の議論に必要な、最小限の原理とアルゴリズムを述べておく。

GMDHで用いられている自己組織化法は、与えられた限られた入出力情報をもとにして、入力変数のランダムな結合を発生させ、その結果生まれる数多くのシステムの中から外部環境に対する適合性に基づいて最良のものを選択するという過程を繰返し、自動的に外部環境に適合するシステムを形成してゆく手法である。

GMDHには、この自己組織化に加えて、種々のヒューリスティクスを導入することにより、大規模で複雑な問題に対しても、比較的容易に最適なシステムの表現式が得られるという一般性がある。主なヒューリスティクスは、

- (i) 入力信号を造るための初等的アルゴリズム
- (ii) 自己選択のための規範
- (iii) 組み合わせの法則

[†] 第13回計測自動制御学会学術講演会で発表(昭49・8)
本研究は文部省特定研究「環境汚染制御」No. 910908に関連して行われたものである。

* 大阪大学工学部 吹田市市山田上

* Faculty of Engineering, Osaka University, Suita
(Received January 25, 1975)

である:

(i)には、たとえばシステムがダイナミカルな場合、そのシステムの実線形なトレンドを算出し、偏差を入力信号として用いるものがある。(ii)には、平均 2 乗誤差、相関係数などがよく用いられる。(iii)としては、2 変数の 2 次多項式、調和関数、直交チェビシェフ多項式、Bayes の式などが用いられる。

GMDH のアルゴリズムはシステムの完全表現式を何に選ぶかによって異なる。完全表現式としてたとえば Kolmogorov-Gabor の多項式

$$\phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \dots \quad (1)$$

があるが、これは 2 変数の 2 次多項式

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2 \quad (2)$$

を部分表現に用いることによって構成することができる。部分表現式に 2 次多項式を用いたアルゴリズムを Fig. 1 に示す。

GMDH の構造は、パーセプトロンと同じく多層構造をしており、(2)式の 2 次多項式を用いた場合、第 T 層を通過すれば、もとの入力変数に関して 2^T 次の多項式となっている。(2)式の係数を決めるには、6 個以上のデータがあればよいが、一般に同じ 2^T 次の多項式を(1)式からただちに求めようとすれば非常に多くのデータが必要であり、係数を決めるために解かねばならない Gauss の正規方程式の次元も大きくなる。

たとえば、入力変数が 10 個のときの完全表現式の項数は約 2 万個あり、(1)式の係数 a_n を決定するには少なくとも同数のデータが必要である。しかも正規方程式に現われる行列の次元は 2 万 $\times$ 2 万となり、計算機の記憶容量をはるかに越える。この意味で、GMDH は少ない入出力データで、高次の多項式近似が可能であり、予測式の構造すなわち入出力関係を学習するための補間点 (interpolation point) の数が完全表現式の項数よりも少ない場合でも最適な近似式が得られる。

GMDH によるモデリングの精度を上げるために用いられるデータの正則化 (regularization) の方法が数種類ある⁷⁾。たとえば、

- (i) データの最後の点がチェックシーケンスに属するようにトレーニングシーケンスとチェックシーケンスに分ける。
- (ii) データの最後の点がトレーニングシーケンスに属するようにトレーニングシーケンスとチェックシーケンスに分ける。

Fig. 1 Algorithm of GMDH

- (iii) データを分散の大ききの順に並べ替え、トレーニングシーケンスとチェックシーケンスに分ける。

などがある。ここでトレーニングシーケンスとは、(2)式の 2 次多項式の係数を決定するためのデータである。この結果生まれる新しいシステムの、外部環境への適合性を調べるためのデータがチェックシーケンスである。このデータは、関数の overfitting を避ける役目も果している。

Fig. 1 の GMDH のアルゴリズムの概要を説明しておく。

(i) 予測式に関与すると思われるすべての変数の中から、相関係数などを使って予測に用いる変数、すなわち入力変数を自己選択する。

(ii) 入力変数の任意の変数の結合 ((2)式) で出力を表わし、トレーニングシーケンスのデータを用いて、Gauss の正規方程式を解き最適係数値を決める。組み合わせ数は nC_2 となる。

(iii) この nC_2 個の関数を、チェックシーケンスのデータに適用し、2 乗誤差を計算して、しきい値より小さなものを若干残しほかはすてる。

(iv) このしきい値を通過した m 個の関数を中間変数とみなし、(ii)、(iii)のステップを繰返す。

(v) 計算停止規範: (ii)、(iii)の繰返し操作を進めてゆくと、2 乗誤差が前層よりも増大しはじめる。このとき計算を打ち切り、最小の 2 乗誤差を与える関数を採用する。

(iv) 各しきい値 θ_i を変え、(v)の最終結果が最適となるように θ_i を設定する。

3. 広域系パターンの同定

システムの信号および信号を伝達する媒質が空間的に連続に分布している広域系は分布定数系と考えられる。このようなシステムの特性は偏微分方程式で記述され、集中定数系に比べて、境界条件の取扱い方、計算の複雑さなどの点において問題の困難さがある。しかも、きわめて多くの因子が複雑に結びついている広

Fig. 2
Example
of the
output of
a large-
spatial
pattern

Table 1 Input data

Interpolation variables	points (2,2)	(3,2)	(3,3)	(4,3)	(3,4)	(2,4)	(2,5)
ϕ_c	ϕ_{22}	ϕ_{32}	ϕ_{33}^*	ϕ_{43}	ϕ_{34}	ϕ_{24}	ϕ_{25}
v_1	ϕ_{12}	ϕ_{22}	ϕ_{23}^*	ϕ_{33}	ϕ_{24}	ϕ_{14}	ϕ_{15}
v_2	ϕ_{21}	ϕ_{31}	ϕ_{32}	ϕ_{42}	ϕ_{33}	ϕ_{23}^*	ϕ_{24}
v_3	ϕ_{32}	ϕ_{42}	ϕ_{43}	ϕ_{53}	ϕ_{44}	ϕ_{34}	ϕ_{35}
v_4	ϕ_{23}^*	ϕ_{33}	ϕ_{34}	ϕ_{44}	ϕ_{35}	ϕ_{25}	ϕ_{26}
$v_5 (=x)$	2	3	3	4	3	2	2
$v_6 (=y)$	2	2	3	3	4	4	5

域系では、偏微分方程式に基づくモデルによって実際の現象をすべて解明することはきわめて困難である。したがってこのような系に GMDH を適用してモデリングを行うことは意義がある。

本論文で作成するシステムモデルは、時間を固定して、システムの出力値や空間的位置のみを使い、任意の位置の出力値を推定するためのモデルであり、過去から現在までの入出力値を利用し、未来の出力値を推定する従来のモデルとは異なる。ここではモデリングの方法としてつぎの2種類を考え、数値例で両者の比較を行う。すなわち、

Type 1: 広域系全体を空間的位置の一つの関数として記述する⁶⁾。

$$\phi = f(x, y) \quad (3)$$

ただし、 ϕ : 広域系の出力値

x, y : 広域系の1点の空間的位置を表わす座標

Type 2: 広域系の1点の出力をその空間的位置およびその点の近傍の出力値の関数として記述する。したがって、予測点が変わると入力と出力の間の関数関係が変化する。

$$\phi_c = f(x, y, \phi_{v1}, \phi_{v2}, \dots, \phi_{vm}) \quad (4)$$

ただし、 ϕ_c : 求めたい広域系の1点の出力値

x, y : その点の空間的位置

Fig. 3
TSS system
structure

$\phi_{vi}, i=1, 2, \dots, m$: その点の近傍の出力値
Type 1 では、一つの予測式(入出力関係(3)式)によって空間全体をモデル化しているのに対して、Type 2 では、予測点ごとに異なった予測式(入出力関係(4)式)すなわち空間全体を複数の関数関係によってモデル化し、予測精度を高めようとしている点に特長がある。

Type 2 の計算方法を簡単に説明する。広域系の出力値が Fig. 2 のように与えられたとき、点(2, 3)の出力値 ϕ_{23} がわからなかったとする。この値を推定するために、この予測点近傍の点を補間点すなわち入出力関係を学習するための点にとると、入力データは Table 1 のように与えられる。ここで $v_i, i=1, 2, \dots, 6$ は入力変数である。 ϕ_{23}^* は ϕ_{23} の近似値で、この値は Type 1 の計算から与えられる。この入力データをもとにして、(4)式の予測式を GMDH によって決定し、点(2, 3)の出力値を予測する。したがって、予測点が異なれば Table 1 の入力データが変わり、予測式すなわち入出力の関数関係も予測点ごとに異なった形で求まってくる。

GMDH を用いて広域パターンを同定して、任意の地点の出力値を求めたい場合、まず最適なモデルを作らなければならない。GMDH では各種しきい値の変更を行う必要があり、ヒューリスティクスの規範を、問題に応じて修正しなければならない。このような状況下では、大型計算機の TSS 機能を用いると、普通のバッチ処理によるよりも、その変更・修正が容易にかつ迅速に行える利点がある。また、パターンの様子をみながら出力値の欲しい地点を指定することも容易にできる。

TSS を用いたシステムの構成(注1)を Fig. 3 に示す。端末から、50 ボー回線制御部を通じて NEAC 2200/500 にデータを送る。各種が Input Command タイプライタに出力されるようにプログラムしておき、それをみて人間が要求にあった条件をタイプなどから入力してやる。FORTRAN で処理された結果

(注1) 大阪大学の TSS システムの一構成図である。

は、再び回線制御部を通じて端末に送り返されてくる。端末には、ミニコンピュータ NEAC-M 4 が接続されていて、ここからのコントロールにしたがって、計算結果がタイプライタなどに出力される。したがって、計算されて出力された結果をみながら、人間の直感によってしきい値を変更して計算をやり直すことができる。

4. 数値例

広域系の出力値として、拡散系のシミュレータによって模擬して得た値を用いる。座標系は、原点を左上隅にとり、下方向に X 軸正方向、右方向に Y 軸正方向をとる。

GMDH の入力データの正規化の方法としては、ここでは分散を計算し、大きなものをトレーニングシーケンスへ、小さなものをチェックシーケンスへ分配する。また、各層で自己選択する中間変数の数はどの層においても一定で、最初の入力変数の数と等しくするものとする。

ここでは、異なった二つのパターンに対して、3章に示した Type 1 と Type 2 の方法を適用し、その精度を比較する。

[例 1]

Fig. 4(a) に示すように、広域系の中央 (4, 4) に常時大きさ 1 の入力があり、これが単位時間経過するたびに四方の格子点に 0.2, 0.2, 0.5, 0.1 の割合で拡散しつつけるものとし、この様子を定常状態になるまで繰返して得た出力値が Fig. 4(b) である。このとき (4, 3), (3, 4), (5, 4), (4, 5) の 4 点の出力値が不明であり、これを求めたい場合を考える。

a) Type 1 を使用した場合

(i) 入力変数: 点 (x, y) における出力値を予測するのにな、 x, y の値を加えて、 x^2, y^2, x^3, y^3 の値を用いる。

(ii) 補間点として Fig. 5(a) に示す 21 点のデータを用いる。

(iii) トレーニングシーケンス/チェックシーケンス=15/6

(iv) 中間変数の自己選択: 中間変数の値と出力値の 2 乗誤差に基づいて、その小さなものから順に 6 個を選択する。

このときの計算結果が Table 2(a) である。等高線の密な点 (3, 4) での値、および入力の影響の強い点 (5, 4) での値の予測は少し悪くなっていることがわかる。このときの最適な予測式は第 3 層で得られた。

b) Type 2 を適用した場合

Fig. 4

Input and output of a large-spatial pattern-example 1
(a) intensity of the input and distribution rate
(b) steady-state output

(i) 入力変

数: 点 (x, y) の出力値を予測する場合、 x, y, x^2, y^2, x^3, y^3 に加え、その点に隣接した 8 点の出力値 $\phi_{x-1, y}, \phi_{x-1, y-1}, \phi_{x-1, y+1}, \phi_{x, y-1}, \phi_{x, y+1}, \phi_{x+1, y-1}, \phi_{x+1, y}, \phi_{x+1, y+1}$ を用いる。

Fig. 5 Arrangement of interpolation points for example 1
(a) Type 1 (b) Type 2

(ii) 補間点として Fig. 5(b) に示す予測点近傍の 11 点を用いる。

(iii) 入力変数の自己選択のためのヒューリスティクス: Type 1 の予測式に含まれる入力変数の中で重要なものを 5 個を選び、それに出力値と $\phi_{x+i, y+j}$ ($i, j = \pm 1, 0$) との相関係数の大きなものを 5 個を選び、この 10 個を用いる。

(iv) トレーニングシーケンス/チェックシー

ケース=6/5

- (v) 中間変数の自己選択：中間変数と出力値との 2 乗誤差に基づいて、その小さなものから順に 10 個選択する。

このときの計算結果が Table 2(b)である。Type 1 で少し精度が悪かった 2 点に対して、精度がよくなっている。ここでは 4 点の予測点に対して、Type 1 で得られた値を出力の近似値として使用している。

Table 2 Computational results for example 1
(a) Type 1 (b) Type 2

Prediction point	Predicted value	True value	Relative prediction error (%)
(4,3)	.371	.382	2.9
(3,4)	.263	.184	42.9
(5,4)	.609	.918	33.7
(4,5)	.323	.382	15.4

(a)

Prediction point	Predicted value	True value	Relative prediction error (%)	Selection layers
(4,3)	.413	.382	8.1	1
(3,4)	.122	.184	33.7	2
(5,4)	.767	.918	16.4	1
(4,5)	.453	.382	18.6	1

(b)

Fig. 6 Input and output of a large-spatial pattern—example 2
(a) intensity of the input and distribution rate
(b) steady-state output

[例 2]

Fig. 6(a)に示すように、広域系の 2 点 (3, 3) と (4, 5) に常時大きさが 2 と 1 の入力はいっている場合について、例 1 と同じくシミュレータで模擬すると、その出力値が Fig. 6(b) のように得られる。予測点として、(3, 3), (3, 5), (5, 3), (5, 5) の 4 点を考える。

a) Type 1 を適用した場合

- (i) 入力変数：点 (x, y) での出力値を予測するに、 x, y の値に加えて、 x^2, y^2, xy, x^3, y^3 の値を用いる。
- (ii) 補間点として Fig. 7(a) に示す 21 点のデータを用いる。
- (iii) トレーニングシーケンス/チェックングシーケンス=12/9
- (iv) 中間変数の自己選択：中間変数と出力値との 2 乗誤差に基づいて、その小さなものから順に 7 個選択する。

このときの計算結果が Table 3(a) である。4 点とも精度が悪いが、特に等高線の密なところの 2 点 (3, 3), (3, 5) の精度が悪くなっている。このときの最適な予測式は第 3 層で得られた。

b) Type 2 を適用した場合

- (i) 入力変数：点 (x, y) の出力値を予測する場合、 $x, y, x^2, y^2, xy, x^3, y^3$ に加え、その点に隣接した 8 点の出力値 $\phi_{x-1, y}, \phi_{x-1, y-1}, \phi_{x-1, y+1}, \phi_{x, y-1}, \phi_{x, y+1}, \phi_{x+1, y-1}, \phi_{x+1, y}, \phi_{x+1, y+1}$ を用いる。
- (ii) 補間点として Fig. 7(b) に示す予測点近傍の 10 点のデータを用いる。
- (iii) 入力変数の自己選択のためのヒュリスティックス：例 1 の b) (iii) と同じ。
- (iv) トレーニングシーケンス/チェックングシー

Fig. 7 Arrangement of interpolation points for example 2
(a) Type 1
(b) Type 2

Table 3 Computational results for example 2
(a) Type 1 (b) Type 2

Prediction point	Predicted value	True value	Relative Prediction error (%)
(3, 3)	.798	.405	97.0
(3, 5)	.468	.259	80.7
(5, 3)	1.767	2.023	12.7
(5, 5)	1.074	1.296	17.0

(a)

Prediction point	Predicted value	True value	Relative prediction error (%)	Selection layers
(3, 3)	.279	.405	31.1	1
(3, 5)	.212	.259	18.1	1
(5, 3)	2.008	2.023	0.7	1
(5, 5)	1.411	1.296	8.9	4

(b)

ケース=7/3

(v) 中間変数の自己選択: 例1の (b) (v) と同じ。

このときの計算結果が Table 3(b) である。Type 1 よりも非常に真値に近い値を得ていることがわかる。ここでも Type 1 で得られた値を出力の近似値として使用している。

例 1, 2.において, 各層で選択する中間変数の数は, 第1層で選択される入力変数の数 (例 1, 2 とともに Type 2 では 10 個) と同じ値にとり, 終始一定値にすえておいている。使用している計算機 (NEAC 2200/700) のコアメモリの容量から, この程度が限度である。また, トレーニングシーケンス/チェックングシーケンスの比は, チェックングシーケンスに対する自乗平均誤差を評価して選択している。ただし, チェックングシーケンスの数は 3 以上としている。Ivakhnenko ら⁸⁾の経験によると, 補間点の総数に対するチェックングシーケンスの割合は 30~40% 程度が適当であるとされており, 例 1, 2 でもそれにならっている。例 2 の Type 2 では, トレーニングシーケンス/チェックングシーケンスの比を 7/3 に選んでおり, このとき予測点 (5, 3) のチェックングシーケンスに対する 2 乗平均誤差は 0.101, 相対予測誤差は 0.7% であるが, いまもしトレーニングシーケンス/チェックングシーケンスの比を 6/4 に変更すると, チェックングシーケンスに対する 2 乗平均誤差は 0.263 に増え, 相対予測誤差は 34.1% (真値 2.023 に対する予測値が 1.334) となる。

以上二つの数値例からもわかるように, Type 1 によって広域系全体を同定し, 任意の点の出力値を近似し, その値も使用しながら, 今度は Type 2 で 1 点

ずつモデリングを行い, 予測点ごとに異なった予測式すなわち入出力関係を求めることによって推定精度を高めるという方法をとると, Ivakhnenko らのアプローチの類系である Type 1 のみを用いるよりも予測精度が向上することがわかる。

5. むすび

広域パターンすなわち時間的には静的な空間的パターンを同定するための方法論として, 独立変数に空間座標をとった GMDH を検討した。さらに, GMDH の中に含まれる種々のヒューリスティクスおよびしきい値を容易にかつ迅速に変更するために, TSS を利用して対話形式で処理する方法を検討した。また, Ivakhnenko らの同定方法 Type 1 に対して, Type 2 と呼ばれる予測点ごとに異なった予測式を求める方法を検討し, その有効性を確めた。空間的なパターンが簡単な因果関係によって生成されている場合には, Type 1 のように入出力の間の一つの関係によって空間全体をモデル化することも可能であろうが, たとえば大気汚染濃度の空間的パターンのように多くの因子が複雑にからみあった結果生成されたパターンを同定するには, Type 2 のように小域的に入出力の間の関数関係を同定し, 複数個 (予測点の数) の予測式によって空間全体をモデル化する Type 2 の方法が有力であるものと思われる。

GMDH はヒューリスティックな要素を多く含んでおり, 与えられた問題に応じて, その問題に適した入力変数, 部分表現式, データの正則化の方法, しきい値の最適化の方法などを決めてやる必要があるが, 既存の手法では手のつけられない複雑なシステムの同定・予測に有効に使える。

本論文では, システムの構造がまったく未知の問題に GMDH を適用しているが, システムの構造がある程度わかっている問題に対して, この情報 (prior knowledge) を有効に活用した上で GMDH をいかに並用してゆくかが今後の課題であろう。

終りに, 「環境制御の方法論」に関するご懇篤なご指導と, GMDH に関する有益なご討論をいただいた京都大学工学部樺本義一教授および池田三郎博士に深甚の謝意を表す。

参考文献

- 1) A. G. Ivakhnenko: Polynomial Theory of Complex Systems, IEEE Trans. on System, Man and Cybernetics, **SMC-1**, 364/378 (1971)
- 2) A. G. Ivakhnenko: Heuristic Self-Organization in Problems of Engineering Cybernetics, Automatica, **6**, 207/217 (1970)

- 3) 池田三郎, 榎木義一: 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, 185/195 (1975)
 - 4) 池田三郎, 藤重 悟, 榎木義一: 河川の流量予測への GMDH の応用, システム制御におけるモデリングとシミュレーションシンポジウム, 59/62 (1974)
 - 5) 井原二郎: 改良形 GMDH—動的な世界人口モデルの場合—, システムと制御, **19**, 201/210 (1975)
 - 6) A.G. Ivakhnenko, et al.: Group Handling of Data in Identification of the Static Cybernetic of a Multi-Extremal Plant, Soviet Automatic Control, **14**, 30/37 (1969)
 - 7) A.G. Ivakhnenko and M.M. Todua: Statistical Prediction of Random Processes Using Self-Organization of the Prediction Equations, Part 1, Problems of Simple Differential Prediction, Soviet Automatic Control, **5**, 35/52 (1972)
 - 8) A.G. Ivakhnenko and P.I. Koval'chuk: Unique Construction of Regression Curve Using Small Number of Points, Soviet Automatic Control, **5**, 26/32 (1972)
-