人　工　知　能　105－1
（1996．8．26）

GAを用いた非線形モデル構築の最適化
—GAとGMDHの融合—

※吉原　郁夫　※※佐藤 周一
※日立製作所 システム開発研究所　※※日立システムテクノロジー

ＧＭＤＨ(General Method of Data Handling)は、非線形モデルを系統的に構築するための有力な手法であるが、モデル構築の際、人間が介在し判断する必要があるため、属人的であったり、時間がかかるという問題があった。

本研究は、遺伝的アルゴリズム(GA)を用いて、有効な説明変数の抽出、および選び出された説明変数の組み合わせ最適化を図ることにより、これら問題点を回避しようとするものである。われわれの試みは、データ数が十分でなかったり、データが誤差を含んでいるなどの現実問題に対処するため、説明変数の取捨選択を明示的に行えるようにした点と、非線形モデルの次数がむやみに高くならないようしている点に工夫がある。

Nonlinear Model Building Method with GA and GMDH

※Ikuo YOSHIHARA　※※Shuichi SATO
※Systems Development Laboratory, Hitachi Ltd.　※※Hitachi System Technology Ltd.

We propose a method to build a nonlinear model with genetic algorithm(GA) and group method of data handling(GMDH). GMDH is widely applicable to many real world problems e.g. time series prediction, system modelling and system identification, but the results depend on the system analyst. Uusing GA makes it possible automatically to build a complex nonlinear model. Our method involves two attempts. One is to limit the degree of the model and the other is to deduce meaningless variables.

1．はじめに

自然や人間社会の複雑な現象及び諸問題を解決するため、さまざまな数理的手法が開発されてきた。対象を数理的に扱うには、何らかの数式モデルを構築する必要があるが、モデル構築には①前提とする理論や法則の欠如、②パラメータ値が不明確、③数量化されない要素の存在などの困難がある、と言われている[1)]。1968年 Ivakhnenko,A.G.が考案したGMDH(Group Method of Data Handling)は、簡単な非線形式を組み合わせて複雑な非線形モデルを自己組織化的に構成してゆく手法であり今日まで予測、モデリング、システム同定などに幅広く応用されている[1) 2)]。

GMDHは、専門家の主観的モデリングに代わって、システムへの入出力情報を基にモデリングを行う一般的かつ柔軟な方法を提供してくれる筈であった。その狙い通り、高次モデルを構築する手続きを構成することは出来たが、低次モデルから高次モデルを構成する過程では、依然、人の介在が必要であり、その結果、属人的であったり、時間がかかる等の問題が残されている[3)]。

本研究は、この部分を遺伝的アルゴリズム(Genetic Algorithm ; GA)を利用し解決しようとするものである。即ち、GA の最適値探索機能利用し、種々の説明変数の中から必要なものを選び出し、且つ、それらの組み合わせを最適化する。従来の研究は、高次モデルを如何に構成するかを狙ったもので、必要な変数の選択はあらかじめ人が行っていることを暗黙の前提としていたが、本研究は、その部分も GA に担わせようとするものである[1) 3) 5)]。

2. GMDH

一口に GMDH と言っても、低次モデルの種類、その低次モデルに入れる変数の取り方などにより、さまざまなバリエーションがある。

最も簡単な低次モデルは2変数の2次式 $G(x_1,x_2)$であり、この式はよく使われている。

$$G(x_1,x_2)=a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 + a_4x_1^2 + a_5x_2^2 \qquad (1)$$

以下では、この中間項 $G(x_1,x_2)$を改めて説明変数とみなし、これらの組み合わせだけで、さらに高次の項を構成する単純な方法を例に、GMDH の概要を説明する(図1)[1)]。

システムへの入力(説明変数)を $x_1,x_2,...,x_m$、出力(被予測変数)を y とする。まず、m 個の説明変数の二つずつの組み合わせを考え、それぞれに対し y を最もよく近似するように関数 $G(x_1,x_2)$ の係数 $a_0,a_1,...,a_5$ を決定する。近似度の良さを測る評価基準は、モデル製作者がそのシステムの目的を踏まえて、経験や知識に基づいて作成する必要がある[3)]。ここに属人的要素が入り込む余地がある。変数の二つずつの組み合わせ $u_{ij}=G(x_i, x_j)$ (ただし、i, j=1,2,...,m ; i<j) は ${}_mC_2$ 通りあるが、その中から出力 y の近似度の良いものから順に p 個を選択する。簡単のため、この p 個を改めて $u_1,u_2,...,u_p$ と記す。

次に選択した $u_1,u_2,...,u_p$ を改めて説明変数とみなし、再び出力 y をよく近似するような組み

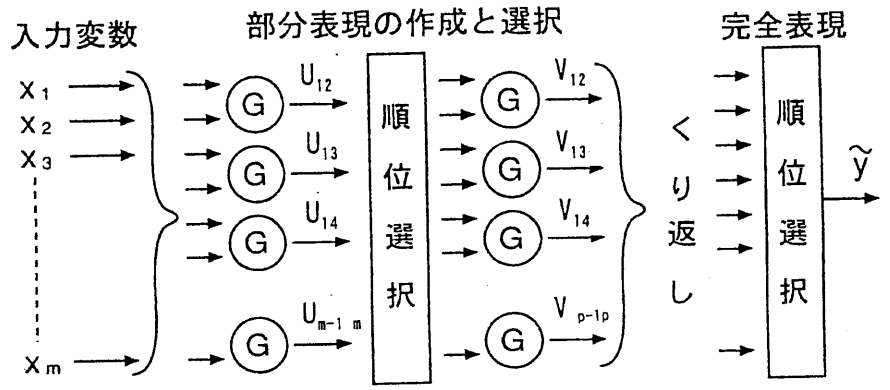


図1　GMDHの概要

合わせ $v_{ij}=G(u_i, u_j)$（ただし i, j=1,2,...,p ; i<j）を探す。はじめと同様にこの中から上位 q 個を選ぶ。以下同様にして、それ以上高次の組み合わせを作っても、y の近似度が上がらなくなるまで、同じ操作を繰り返す。このようにして求められた中間段階のモデル式 u_{ij} や、v_{ij} を「部分表現」と呼び、最終のものを「完全表現」と呼ぶ。

なお一般には、部分表現同士を組み合わせて次の部分表現を作るとは限らない。遺伝的プログラミング(GP)を使う場合、部分表現または説明変数を組み合わせ、次の部分表現(高次モデル)を構成して行くのが普通である[5]。この場合、モデルの構成は、図1のようなネットワークよりも木構造で表す方が自然である。

3. GA と GMDH を用いる非線形モデル構築

GMDH の部分表現の選択に GA を用いる。

3.1 モデル構成の二分木表現

低次モデル式としては(1)式と同じ G(·,·)を用い、説明変数から部分表現を作り、その部分表現同士を組み合わせて次の部分表現を作って行く。この際、説明変数または部分表現は重複して使わないこととする。このような制約を課すと、モデルの次数が低く抑えられるが、次数を抑える理由は「多くの実応用においては、データ数が不十分であったり、データが誤差を含むため、線形では不十分だとしてもあまり高次にしない方が良い」との経験を反映したいからである。

上記モデル構成は、説明変数を葉とし、各 G(·,·)をノードとする二分木で表現できる。木の高さは $\underline{h}=\log_2 n\uparrow$ (n は説明変数の数、↑は切り上げの意)である。$\underline{h}$ は、G(·,·)を用いて合成してできる木の高さの最小値である。逆に、高さ $\underline{h}$ の二分木が表せるのは、$n^*/2<n\leqq n^*$(ただし $n^*=2^{\underline{h}}$)個の説明変数のモデルであり、$n<n^*$ の場合は葉の一部が欠けた木となる。そこで、説明変数 $\mathbf{X}=(x_1,x_2,...,x_n)$に対し、ダミー変数 d を n^*-n 個加えて、$\mathbf{X}'=(x'_1,x'_2,...,x'_{n^*})$とし、常に完全な二分木となるようにする。このようにすると、d と x_i の組み合わせが生じるが、これは単に x_i そのものとして扱うことにする。

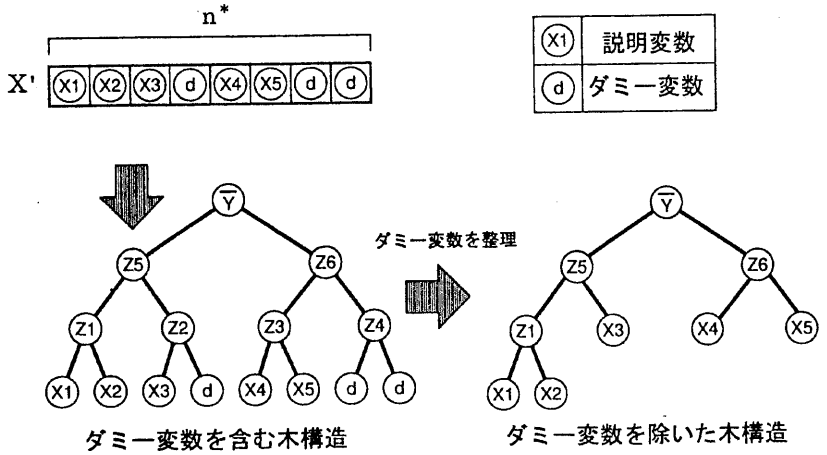


図2　GMDH モデル構成の二分木表現

3.2 GA による説明変数の組合せ最適化

X'を構成するn*個の要素（x_i及びd）の並びをそのまま染色体とする。変数の並びが遺伝子型であり、合成されたモデルが表現型である。主な遺伝的操作は次の通りである。

a）交差……一点交差の位置cをランダムに決定する。親個体1の染色体からcの左側の遺伝子列をとり、そのまま子個体に転写する。次に親個体1から取らなかった遺伝子を、親個体2から出現順に取り出し、子個体のcの右側に転写する。こうすることにより、致死遺伝子の発生、即ちデータの重複・欠落を防ぐことが出来る。

b）突然変異……染色体の一点をランダムに選び、その右側と左側の遺伝子列を入れ換える。ただし、最優良個体(エリート)に対しては、突然変異は施さない。

c）個体評価と自然選択……各個体は、評価データに対する推定値(GMHD モデルの出力)と真値yの自乗誤差で評価する。誤差が小さい個体ほど適応度が高いと考え、順位選択する。

3.3 GA による説明変数の取捨選択

有効でない説明変数は、係数 $a_0,a_1,...,a_5$ が0に近づくことにより、自ずと消えていく筈である。しかし、データの数が少なかったり、誤差などのため、係数がうまく減少しない場合も生じる。そこで、われわれは不要な説明変数をモデルから明示的に取り除けるようにしようと考えた。遺伝子長n*をさらにIだけ拡張し、長さn*+I とした染色体X"を考える。

$$X''=(x''_1, x''_2,\ldots ,x''_{n^*+I}) \qquad (2)$$

ここで、X"の成分のうちn個はもとの変数x_iであり、n*+I−n個はダミー変数dである。この染色体のうち、左側のn*個の遺伝子だけを使って木を構成し、右側のI個の遺伝子は、木の構成には使わない。染色体X"を組替えることにより、最大I個の説明変数が取り除かれるモデルが、次々に生成される。図3は、説明変数$x_1,x_2,...,x_6$にI=2個のダミー変数を追加した結果、x_6がモデルから取り除かれた例である。

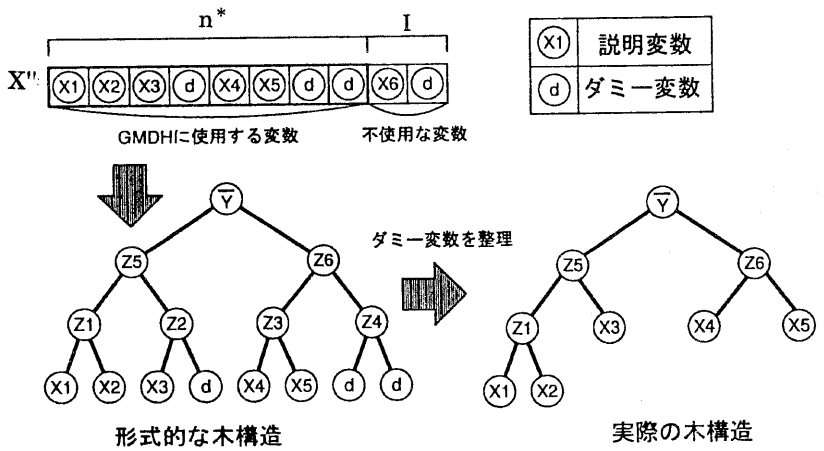
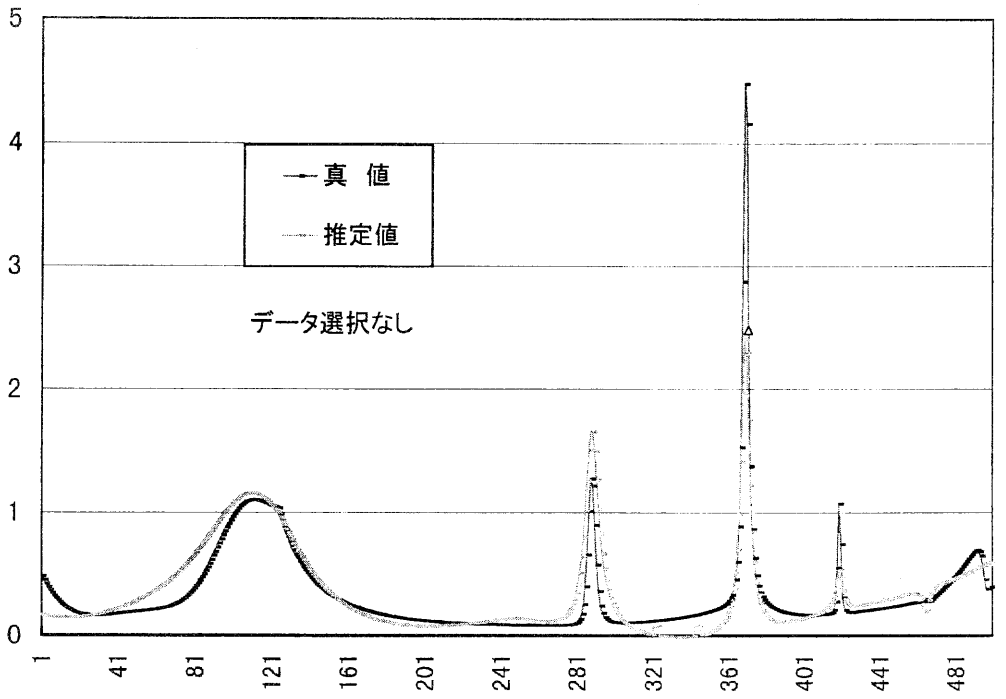


図3　説明変数の取捨選択



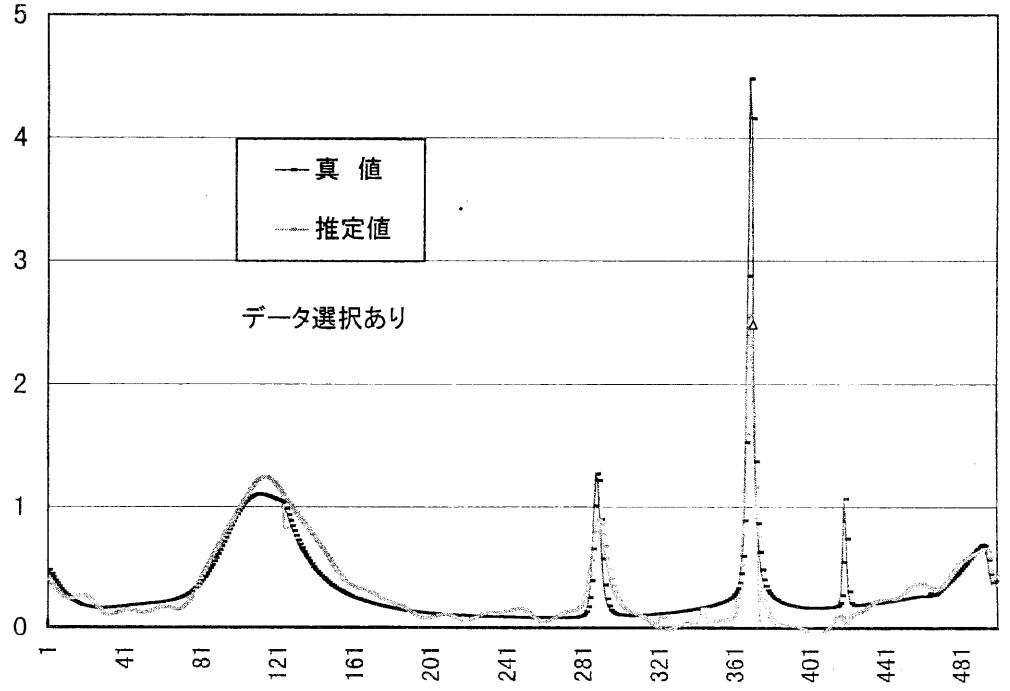


図4　GMDH による推定結果

4．モデル検証

8 つの説明変数からなる多項式 $f(X)= f(x_1,x_2,...,x_8)$により生成したデータを用いて、本手法でどのような推定値が得られるかを見る。一つのデータは 500 点からなり、その 80%をモデル決定に用い、10%を適応度評価(各個体が表すモデルの相互比較)に用いた。残りの 10%は、未知データに対する評価用に残した。個体数 64、交差率 20%、突然変異率 5%とし 300 世代まで計算し、そのときの最優良個体が完全表現を与えると考える。

図 4 上段に、8 つの説明変数だけを与えたときの、推定値を示す。同図の横軸はデータの入力順に対応し、縦軸は真値及び推定値で、スケールは任意である。図は一見、時系列予測のように見えるが実はデータの並び順に特別の意味はない。各点で高さがどれだけ一致しているかだけが意味をもつ。複雑な波形の山谷はおよそ一致していることから、非線形モデルを構築した効果は見られるが、誤差はかなりある。

次に、無関係の変数 4 を加え、説明変数候補を 12 とし、変数の選択及びモデル構築を行ってみる。図 4 下段は、変数選択も行わせたときの結果である。この場合の方が一般には難しい筈であるが、ピーク位置などはほぼ正しく推定している。図で左側のなだらかな山の付近は、上図よりもむしろ合っているが、右側のピーク一つを逃し、標準偏差でみた推定誤差は 2 倍以上大きかった。真の変数全部は見つけられないことの影響と考えられる。

このような人工データ 10 ケースで検討したところ、変数の選択は一部誤っているが、推定誤差はむしろ少ないこともある。別途おこなった実データの解析では、推定精度は十分ではないが、モデルに組み込まれた変数に妥当性があり、主要な説明変数の抽出に有益であることがわかった。

5.おわりに

GMDH と GA を融合した、非線形モデル自動構築法を提案した。これにより、モデル構築の途中段階においての経験判断の介入が不要となり、探索の偏りを減らすこと及びシステム構築時の労力を軽減することができる。

説明能力の低い変数をモデルに組み込まないようにするための遺伝子コーディング法が一つの工夫点であり、これにより、あらかじめ多くの説明変数の中から重要なものを人が選り分けるという事前準備作業も軽減できると考えている。

これまでの検討の結果では、低次モデルから高次モデルを構成する手続きが確定的であるという、GMDH が持っていた限界が現れているように思える。ＧＡの特徴である大域探索機能を生かした改良を図りたい。

<参考文献>

1）池田：GMDH(変数組合せ計算法)の基礎と応用:システムと制御、Vol.23, No.12, pp.710-717, 1979

2）Ivakhnenko,A.G. : The Group Method of Data Handling, A Rival of the Method of Stochastic Approximation : Soviet Automatic Control, Vol.1, pp.43-55, 1968

3）Tenorio and Ree:Self-Organizing Network for Optimum Supervised Learning:IEEE Tr.Neural Networks, Vol.1, No.1, pp.100-110, 1990

4）吉原、佐藤：GA による GMDH モデル構築の最適化、情処第 53 回全大、3B-11,1996

5）Iba,H, Garis,H., Sato,T : A Numerical Approach to Genetic Programming for System Identification Evolutionary: Computation, Vol.3, No.4, pp.417-452, 1995