第 20 卷 第 4 期　　自 动 化 学 报　　Vol. 20, No. 4
1994 年 7 月　　ACTA AUTOMATICA SINICA　　July, 1994

短 文

GMDH 中部分表达式的构成及改进方法

徐田军　王桂增
（清华大学自动化系　北京　100084）

摘　要

该文在综合了 GMDH 方法中部分表达式构成形式的基础上，提出用最优化方法构成系统最佳部分表达式，由此得到精度更高、稳定性较好的模型。

关键词：GMDH 部分表达式，最优化方法，非线性系统辨识。

1　引言

GMDH 方法（Group Method of Data Handling）是前苏联乌克兰科学院 A.G. Ivaknenko 运用多层神经网络原理和品种改良假说，以 K-G 多项式为基础提出的一种复杂非线性系统的辨识方法——数据处理组合方法，简称 GMDH 方法[1]。

考虑一非线性系统

$$y = f(x_1, x_2, \cdots, x_n) \qquad (1)$$

式中，$x_1, x_2, \cdots, x_n$ 是输入，y 是输出。

函数 f 的离散 Volterra 级数展开式（Kolmogorov-Gabor 多项式，简称 *K-G* 多项式）

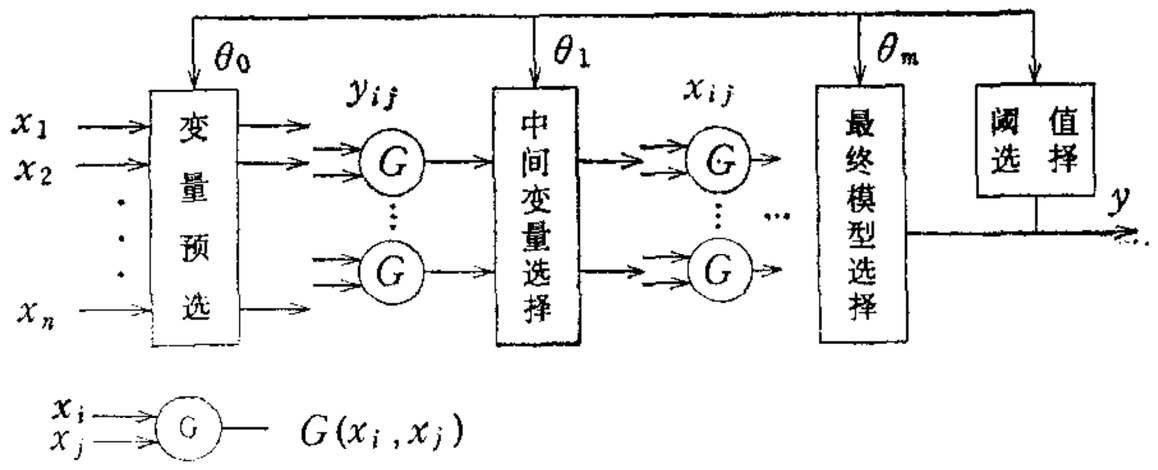

图 1　基本型 GMDH 方法示意图

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \cdots \qquad (2)$$

被广泛地用来作为非线性模型的完全描述。但要完全确定 $a_0, a_i, a_{ij}, \cdots$ 等参数的值是不现实的，因为当 n 比较大时，会产生巨大的维数灾。GMDH 方法主要是利用多层叠代的方法，逐层选择，以获得系统的最终模型。基本型 GMDH 方法示意图如下：

图中 G 称为部分多项式，它是两个输入变量的完全多项式，θ_1、θ_2、$\cdots$、θ_m 分别为每层的模型评价准则值，y_{ij} 是由部分模型计算得到的输出，部分模型是通过拟合实测数据

本文于 1992 年 12 月 2 日收到

辨识得到的；x_{ij} 称为中间变量，是从 y_{ij} 中按每层的模型评价准则值选出来的，作为下一层的输入．假定通过 m 层得到最终模型，且假定部分多项式都是二阶的，则最终模型是 2^m 阶结构的多项式．最终模型的质量不仅与中间变量的选择有关，更与构成中间变量的部分表达式的形式有关．

2　部分表达式的构成形式

在 GMDH 方法中存在大量需要凭借经验予以解决的问题，其中部分表达式的构成决定了最终模型的基本形式．部分表达式的构成一般有以下几种形式．

1）多项式形式．

这是在 GMDH 方法中使用最多的一种形式，这种形式的基本函数为：

$$y = a_0 + a_1x_i + a_2x_j + a_3x_i^2 + a_4x_j^2 + a_5x_ix_j \tag{3}$$

$$y = a_0 + a_1x_i + a_2x_j + a_3x_ix_j \tag{4}$$

$$y = a_0 + a_1x_i + a_2x_j \tag{5}$$

基本 GMDH 方法的创始人 A.G. Ivaknenko 提出的是(3)式所示的完全二元二次多项式[1]，以致模型最终表达式阶次比较高，形式复杂，不稳定；且高层随中间变量间的相关性增加，线性方程组病态愈加严重，计算发生困难．

田村坦之的改进型 GMDH 方法是对基本型 GMDH 方法的重大改进[2]．这种方法是在(1)式的系数 a_1、a_2、$\cdots$、a_5 中，任意令其中 $n(0 \leqslant n \leqslant 4)$ 个系数为零，从得到的 $C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5 = 31$ 种不同的组合结构式中选一最佳结构式，参加中间变量选择．

在数据较少时为了增加模型的稳定性，井原提出采用带滤波作用的二次型函数作为部分表达式[3]，如下所示：

$$f(x_i, x_j) = a_0 + a_1z_{ij} + a_2z_{ij}^2$$

$$z_{ij} = wx_i + (1 - w)x_j \tag{6}$$

式中 w 是权系数，$0 < w < 1$．

以上这些多项式形式表示的部分表达式形式比较简单，且适用范围广、参数处理方便．

2）有理函数形式．

在采用以上形式不能得到满意的结果时，D.L. Dylbokova 提出采用如下有理函数形式：

$$f(x_i, x_j) = \frac{a_0 + a_1x_i + a_2x_j + a_3x_i^2 + a_4x_j^2 + a_5x_ix_j}{1 + b_1x_i + b_2x_j + b_3x_ix_j} \tag{7}$$

采用(7)式所示的部分表达式需要用非线性最小二乘法辨识参数，无疑运算量将大大增加．

3）逻辑函数形式．

适用于输入、输出仅取二值的系统．

4）概率密度函数形式．

适用于统计决策中，求某一系统的完全表达式（Bayes 公式），即非参数辨识预报.

5）逐次导人型.

适用于需要求解微分方程、偏微分方程的领域.

6）非线性指数型.

适用于一些用此种函数能更好地描述的系统，如自然环境模型、农作物生长模型等.

7）样条函数形式.

R. Mehra 认为，部分表达式是 GMDH 型方法的薄弱环节[4]. 采用（3）式难以适合不同的系统，而采用(7)式又要求解非线性方程. 他建议采用插值的方法，用样条函数作为部分表达式.

3 改进部分表达式构成的新方法

在以上七种部分表达式形式中，多项式形式最是常用的，但它仍有许多不尽人意的地方，难以适合不同的系统. 特别是在中间变量保留数目受限制的情况下，难以实现充分的自组织，影响系统最终模型的质量. 因为中间变量保留数目多，会使运算工作量大，并占用较多的计算机内存；但中间变量保留数目过小，又会过早淘汰有用变量，使最终的模型质量不高. 杨自厚提出的原始变量保存算法[4]，可以较好地解决这一矛盾，这种方法改变了基本 GMDH 方法中的变量传输方法，在各层均设立一个原始变量保存区，将作用较弱的变量保存起来，直接参与各层的组合运算，避免了在低层过早被淘汰. 这样就可以使自组织更加充分，并为改进部分表达式的形式及中间结果提供所需的信息.

原始变量保存，一般选择那些可控变量以及机理上认为比较重要的变量.

正如 R. Mehra 指出的那样，采用(3)式难以适合不同的系统，而采用(7)式又需要解非线性方程. 作者提出的变量优化算法既可把(3)、(7)式有效地结合起来，拓宽部分表达式的描述范围；又可以避免解复杂的非线性方程. 这种方法的基本思想是通过相关系数分析获得最佳"预报因子"，以此代替原始变量参与部分表达式的构造.

在回归分析中，相关系数是用来描述变量间关系密切程度的一种数量指标. 相关系数有简单相关系数和偏相关系数. 如果能在一类初等函数的集合中，利用相关系数分析方法，选取适合预报对象的函数形式作为"预报因子"来代替原始变量参加部分表达式的构成，则能得到多种形式的部分表达式，并可提高预报精度. 考虑两变量 x_i 和 x_j，设它们的"预报因子"的函数形式分别为 $f_i(x_i)$ 和 $f_j(x_j)$，则 $f_i(x_i)$ 与 $f_j(x_j)$，$f_i(x_i)$ 与 y，$f_j(x_j)$ 与 y 之间的简单相关系数分别为

$$r_{ij}=r_{ji}=\frac{\Sigma[f_i(x_{it})-\overline{f_i(x_{it})}]\,[f_j(x_{jt})-\overline{f_j(x_{jt})}]}{\sqrt{\Sigma\,[f_i(x_{it})-\overline{f_i(x_{it})}]^2\Sigma\,[f_j(x_{jt})-\overline{f_j(x_{jt})}]^2}}, \tag{8}$$

$$r_{iy}=r_{yi}=\frac{\Sigma\,[f_i(x_{it})-\overline{f_i(x_{it})}](y_t-\bar{y}_t)}{\sqrt{\Sigma\,[f_i(x_{it})-\overline{f_i(x_{it})}]^2\Sigma(y_t-\bar{y}_t)^2}}, \tag{9}$$

$$r_{jy}=r_{yj}=\frac{\Sigma\,[f_j(x_{jt})-\overline{f_j(x_{jt})}](y_t-\bar{y}_t)}{\sqrt{\Sigma\,[f_j(x_{jt})-\overline{f_j(x_{jt})}]^2\Sigma(y_t-\bar{y}_t)^2}}. \tag{10}$$

式中，

$$\overline{f_i(x_{it})} = \frac{1}{n}\Sigma f_i(x_{it}),\quad \overline{f_j(x_{jt})} = \frac{1}{n}\Sigma f_j(x_{jt}),\ \bar{y}_t = \frac{1}{n}\Sigma y_t$$

Σ 代表 $\sum_{t=1}^{n}$，n 为数据组数。

令

$$A = \begin{vmatrix} r_{ii} & r_{ij} & r_{iy} \\ r_{ji} & r_{jj} & r_{jy} \\ r_{yi} & r_{yj} & r_{yy} \end{vmatrix}, \tag{11}$$

则 y 与 $f_i(x_i)$，y 与 $f_j(x_j)$ 之间的偏相关系数分别为

$$R_{yi} = -A_{iy}/\sqrt{A_{yy}\,A_{ii}},$$

$$R_{yj} = -A_{jy}/\sqrt{A_{yy}\,A_{jj}},$$

其中 A_{iy}，A_{yy}，A_{ii}，A_{jy}，A_{jj} 分别为 A 中 $r_{iy}, r_{yy}, r_{ii}, r_{jy}, r_{jj}$ 的代数余子式。

选择 $f_i(x_i)$ 和 $f_j(x_j)$ 的形式使偏相关系数

$$R_{ij} = R_{yi} + R_{yj} \tag{12}$$

取得最大值。

函数 $f_i(x_i)$ 形式可选择以下几种：

$$\text{①}\quad f_i(x_i) = \frac{1}{x_i},\qquad \text{②}\quad f_i(x_i) = x_i,$$

$$\text{③}\quad f_i(x_i) = x_i^2,\qquad \text{④}\quad f_i(x_i) = x_i^3. \tag{13}$$

每次任选二种函数形式 $f_i(x_i)$ 和 $f_j(x_j)$ 对它们进行相关性分析，从 $C_4^2 = 12$ 种组合中选出偏相关系数最大的一组函数形式作为最佳的“预报因子”，以其代替原变量参加部分表达式的构成。

为了使部分表达式的自组织更加充分，作者借鉴了原始变量保存算法的思想，选择一些在机理上对辨识参数有较大影响的原始变量进行保存，使它们能在高层同最佳“预报因子”进行再组合，这样就可以构成形式多样的有理多项式，大大提高了模型的精度。

对于易于用非线性指数形式描述的系统，作者提出用一类在实际中应用很广的BOX-COX 变换函数族来构成“预报因子”[6]，如下式：

$$f_i(x_i) = (x_i^{a_i} - 1)/a_i,\quad (-\infty < a_i < +\infty), \tag{14}$$

其中 a_i 为变换参数，当 $a_i \to 0$ 时，$f_i(x_i) \to \ln x_i$；当 $a_i = 1$ 时，$f_i(x_i) = x_i - 1$。这类函数包括了对数线性与线性关系，是一类广泛的幂函数。

“预报因子”的函数形式的选择很关键，本文只讨论了部分函数形式。作为该方法的拓宽，可以增加不同的函数形式，以进一步提高应用效果。

4　仿真结果

某石化公司聚丙烯装置是引进日本三井油化本体法生产工艺。其中熔融指数是确定

产品牌号、质量的重要因素，作者应用改进的 GMDH 方法对熔融指数的预报作了一些研究。

从现场操作报表中经过数据整理后取出八个参数的 45 组数据分成两组，前一部分共 30 组数据用于拟合建模，后一部分共 15 组数据用于模拟预报(数据组略去)。

在采用的各种改进的 GMDH 方法中，模型检验准则均采用改进的 AIC 准则。

应用上述三种改进的 GMDH 算法所得到的模型结果如下表，其中，M表示得到最终模型的运算层数，AIC 表示最终模型的 AIC 值。

表 1 三种模型结果比较

	田村坦之改进型 TAMURA GMDH	原始变量保留算法 IVK GMDH	变量优化算法 VO GMDH
M	4	5	4
AIC	−15.5	−24.8	−40.7

由表 1 可以看出，采用变量优化算法，预报精度要高出前二种改进算法。

参 考 文 献

[1] Ivaknenko A G. Heuristic Self-Organization in Problems of Engineering Cybernetics. Automatica, 1970, **6**: 207—219.

[2] Tamura H, Kondo T. Heuristic Free Group Method of Data Handling Algorithm of Generating Optimal Partial Polynomials with Application to Air Pollution Prediction. *Int. J. Systems Sci.*, 1980, **11**: 1095—1111.

[3] Hara K, Yamamoto T, Terads K. Prediction by Dual Mode GMDH. *Int. J. Systems Sci*, 1988, **19**: 2673—2681.

[4] Mehra R K. Nonlinear System Identification Selecteed Survey and Recent Trends. Preprints 5th IFAC Symp. on Identification and Parameter Estimation, 1979. 77—83.

[5] 杨自厚，李宝泽. 改进的GMDH方法——原始变量保存算法. 自动化学报，1986，**12** 397—400.

[6] 徐田军，王桂增.改进的GMDH 方法及其在聚丙烯熔融指数预报中的应用. 过程模型化与控制，第 6 卷，科学出版社，1993.

[7] 钟登华 刘豹.非线性计量经济建模的一种有效方法. 系统工程，1992，**10**: 42—47

THE CONSTRUCTURE OF PARTIAL POLYNOMIALS AND IMPROVED ALGORITHMS IN GMDH

Xu Tianjun　Wang Guizeng

(Department of Automation Tsinghua University)

Abstract

Based on synthesizing the constructures of partial polynomials in GMDH, the optimal constructure of partial polynomials is obtained by using optimization method in this paper. The model obtained is more accurate and stable.

Key words: GMDH, Partial polynomials, Optimization method, Identification of non-linear system.