DOI:10.11721/cqnuj20150414

基于快速数据分组处理方法的改进研究*

王 爱 华
(南充职业技术学院 人文艺术系, 四川 南充 637131)

摘要:数据分组处理方法(GMDH 算法)是一种模拟大脑辨识复杂非线性系统的算法。这种算法在对数据进行分组处理时具有较明显的优势。但是处理大数据时,当取恰当终止拟合条件,GMDH 算法又显得比较弱势。为此从 GMDH 运算的本质去解决算法的效率问题,给出了一种结合快速求解逆矩阵以达到快速求解拟合方程的目的的算法,并且利用外推准则,合理选取拟合结束条件及快速选择最佳模型的方法。在不改变 GMDH 算法准确度的情况下,研究了如何提高 GMDH 算法的计算效率。

关键词:快速数据处理算法;最小二乘法;拟合原理;残差;逆矩阵
中图分类号:TP183;O29　　**文献标志码**:A　　**文章编号**:1672-6693(2015)04-0113-05

1967 年,乌克兰数学家 Ivakhnenko 首次提出数据分组处理方法(Group method of data handling,GMDH 算法)[1],也称感应学习算法,是一种模拟大脑辨识复杂非线性系统的算法,GMDH 算法在很多方面都能较有效地作短期预测,尤其在模式识别、数学建模及对随机过程的预测方面有重要的应用[2-6]。但实践证明,当处理的数据集过于庞大时,处理效率不是很理想;如果数据集过小或者不够全面,又不能达到预期目的。基于这种算法的优越性和不足,很多学者对其进行了更深入地研究,鲁茂和张秋菊对快速 GMDH 算法分别从不同的角度进行了改进并取得了一些成果[7-8],其中主要的研究方向是,如何对 GMDH 算法进行算法的完善和改进? 本文从 GMDH 运算的本质出发,在不改变 GMDH 算法准确度的情况下,研究了提高 GMDH 计算效率的一种新方法——快速求逆矩阵法,这种算法将大大提高 GMDH 的计算效率。在此,首先介绍最小二乘法及最小二乘法拟合原理在 GMDH 算法中的应用。

1 最小二乘法拟合原理与 GMDH 算法

1.1 最小二乘法拟合原理

最小二乘法是根据最小二乘准则原理[9-10],利用样本数据拟合回归方程的一种算法。从整体上寻找一个近似函数 $y(x)$,使得该函数与样本数据$(x_i,y_i)(i=0,1,2,\cdots,n)$的误差 $e_i=y_i-y(x_i)(i=0,1,2,\cdots,n)$最小。常用的方法有 3 种:第 1 种是误差的绝对值的最大值$\max\limits_{0\leqslant i\leqslant n}|e_i|$最小;第 2 种是误差绝对值之和 $\sum\limits_{i=0}^{n}|e_i|$ 最小;第 3 种是误差的平方和 $\sum\limits_{i=0}^{n}e_i^2$ 最小。前两种方法虽然简单、自然,但不便于微分求解,第 3 种方法常常被称之为最小二乘法。在统计学与经济学等学科及应用中,最常见的近似函数是

$$y=\sum_{i=0}^{n}\beta_i x_i+\varepsilon\ 。\tag{1}$$

其向量表达式为 $y=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon$,其中 $X=[1\quad x_1\quad x_2\quad\cdots\quad x_n]^{\mathrm{T}}$,$\boldsymbol{\beta}=[\beta_0\quad\beta_1\quad\cdots\quad\beta_n]^{\mathrm{T}}$,$\varepsilon\sim N(0,1)$。

1.1.1 多元线性回归模型的参数估计　由上述原理可知,最小二乘法(OLS)拟合的实质是寻找一个函数 $y(x)$,

* 收稿日期:2015-01-04　修回日期:2015-03-10　网络出版时间:2015-4-20 09:30
作者简介:王爱华,女,副教授,研究方向为遗传算法与数学教育,E-mail:1246840742@qq.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20150420.0930.005.html

使得其与样本数据的误差 e_i 的平方和最小,即:

$$Q=\sum_{i=0}^{n}e_i^2=\sum_{i=0}^{n}(y_i-y(x_i))^2=\sum_{i=0}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_nx_{in})^2 \tag{2}$$

最小。由多元函数极值的必要条件,得:

$$\frac{\partial Q}{\partial\beta_i}=0,i=0,1,2,\cdots,n。 \tag{3}$$

即:

$$\begin{cases}\dfrac{\partial Q}{\partial\beta_0}=\sum\limits_{i=0}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_nx_{in})\times(-1)=0,\\ \dfrac{\partial Q}{\partial\beta_1}=\sum\limits_{i=0}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_nx_{in})\times(-x_{i1})=0,\\ \vdots\\ \dfrac{\partial Q}{\partial\beta_k}=\sum\limits_{i=0}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_nx_{in})\times(-x_{in})=0。\end{cases} \tag{4}$$

解方程组(4)即可得到参数的估计值 $\hat{\beta}_i(i=0,1,2,\cdots,n)$。

若用矩阵表示,方程组(4)可转化后表示为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$,当 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为可逆矩阵时,有

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{Y}。 \tag{5}$$

1.1.2 可决系数 要定量判断样本观测值与回归方程的拟合程度,而这种定量的判断指标就是可决系数 R,是一种综合度量回归模型对样本观测值拟合优度的度量指标。可决系数越大,说明模型对样本观测值的拟合优度越好;反之可决系数越小,说明模型对样本观测值的拟合优度越差,可决系数计算公式:

$$R^2=\frac{RSS}{TSS}=1-\frac{ESS}{TSS} \tag{6}$$

(6)式中,RSS,TSS,ESS 分别表示残差平方和、回归平方和及总离差。

1.2 最小二乘法在 GMDH 算法中的应用

在 GMDH 算法中应用最小二乘法的大致思路是:设所研究样本的数据容量为 n,将此 n 个数据分成两个集合 n_A 和 n_B,其中 n_A 是训练集,n_B 是测试集,在输入变量与输出变量间建立一个参考函数,在标准的规则中选择合适的目标函数,以参考函数为基础,用最小二乘法拟合的方法求解参数 $\hat{\beta}_i(i=0,1,2,\cdots,n)$,再用测试集 n_B 测试拟合方程的可决系数 R(可由(6)式计算得出),在此基础上就可以选择最好的测量方程作为最佳模型。

由高斯-马尔可夫定理可知,如果上述条件成立,则最小二乘估计值 $\hat{\beta}$ 是 β 的最优线性无偏估计。但是由于上述方法在实际应用中常常需要处理 β_i,其中的 i 值偏大,加上较少的数据经过 GMDH 一次运算后,数据量急剧增加,势必在计算如此庞大数据量上会消耗很多的时间。

因此,为了在不损失精确度的前提下,加速最小二乘法拟合的过程,就应该思考如何改进求解 β_i 的过程,以达到快速求解 β_i 的目的。下面先讨论经典 GMDH 算法。

2 经典 GMDH 算法[11]

对于未知非线性系统,GMDH 算法用离散函数的高阶 K-G 多项式来表达。设:

$$y=f(x_1,x_2,\cdots,x_n)=a_0+\sum_{i=1}^{n}a_ix_i+\sum_{j=1}^{n}\sum_{i=1}^{n}a_{ij}x_ix_j+\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}a_{ikj}x_ix_jx_k+\cdots, \tag{7}$$

GMDH 算法经过多层筛选,用局部简单的参考函数不断组合逼近(7)式,可以得到整体比较复杂的模型。其中,使用的参考函数为二元二次函数:

$$y=f(x_1,x_2)=a_0+\boldsymbol{A}_1\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_2\boldsymbol{X} \tag{8}$$

其中 $\boldsymbol{A}_1=(a_1 \quad a_2)$,$\boldsymbol{A}_2=\begin{pmatrix} a_3 & \frac{a_5}{2} \\ \frac{a_5}{2} & a_4 \end{pmatrix}$,$X=(x_1 \quad x_2)^{\mathrm{T}}$。

用经典 GMDH 算法建立数学模型的具体步骤如下。

1) 样本空间 W 的样本数据为 n 个,将 W 划分为 n_A(训练集)和 n_B(测试集),划分的条件是训练集与测试集为完备互斥的,即 $n_A \bigcup n_B=W$,$n_A \bigcap n_B=\varnothing$;

2) 建立输入输出的参考函数关系式:

$$y_{1k}=f(x_1,x_2)=a_0+\boldsymbol{A}_1\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_2\boldsymbol{X} \tag{9}$$

3) 从第一层开始,在训练集 n_A 上利用内准则建立中间模型。y_{11} 为第一层的第一个中间模型,y_{2k} 为第二层的第 K 个中间模型,以此类推。用最小二乘法估算 y_{1k} 的系数;

4) 在测试集 n_B 上外准则决定中间模型的去留。对 $\{y_{1k}\}$ 进行筛选,选出较优的中间模型 $y_{1j}(j\leqslant k)$ 作为第二层;

5) 重复上述过程,则可产生第二、第三层等等,在哪层产生最优外准则就在哪层停止建模工作,选择最好的测量方程作为最优复杂度模型。

从上述建模过程可以看出,用经典 GMDH 算法建模虽然可以得到精确度较高的模型,且方法简单易掌握,但计算量大,而且得到的模型相当复杂。因此,有必要研究如何降低 GMDH 的复杂度,首先要想办法快速求出 β_i,这里将要介绍的方法是——快速求逆矩阵。

3 快速求逆矩阵

3.1 快速求逆矩阵的步骤

快速求逆矩阵又称为高斯—约旦法求逆矩阵,其计算步骤如下。

首先,对于 k 从 0 到 $n-1$ 作如下几步:

从第 k 行、第 k 列开始的右下角子阵中选取绝对值最大的元素,并记住次元素所在的行号和列号,在通过行交换和列交换将它交换到主元素的位置上,这一步称为全选主元。

$$m(k,k)=1/m(k,k) \tag{10}$$

$$m(k,j)=m(k,j)*m(k,k),j=0,1,\cdots,n-1,j!=k \tag{11}$$

$$m(i,j)=m(i,j)-m(i,k)*m(k,j),i,j=0,1,\cdots,n-1,i,j!=k \tag{12}$$

$$m(i,k)=-m(i,k)*m(k,k),i=0,1,\cdots,n-1,i!=k \tag{13}$$

最后根据全选主元过程中记录的行、列交换的信息进行恢复,恢复的原则如下:在全选主元的过程中,对先交换的行(列)后进行恢复;原来的行(列)交换用列(行)交换来恢复,得到的新矩阵就是逆矩阵 $\boldsymbol{\beta}_i$,这种方法快速、准确、计算量小,非常适用。

3.2 效率对比

基于前面介绍的快速求逆矩阵方法,下面通过表 1 将两种算法进行对比。

表 1　两种算法效率比较

	原算法	原算法(高度优化)	新算法
加法次数	103	61	39
乘法次数	170	116	69
需要的额外空间	$16*sizeof(float)$	$34*sizeof(float)$	$25*sizeof(float)$

由表 1 可以看出,使用新算法大大降低了计算量,提高了计算效率,其优越性是显而易见的。

4 快速选择最佳模型

事实证明,运用快速求逆矩阵的方法的确减少了运算量。那么,如何选择最佳模型? 通常采用 2 种方法——快速选择法和大小堆法。

张宾等人提出,从参考函数构成的初始模型(函数)集合出发,按一定的法则产生新的中间候选模型(遗传、

变异),再经过外准则筛选(选择),重复这样一个遗传、变异、选择和进化的过程,使中间待选模型的复杂度不断增加,直至得到最优复杂度模型[4]。GMDH 算法是数据分组和贯穿于整个建模过程中的内、外准则的运用。它在训练集 n_A 上,利用内准则建立中间待选模型;在测试集 n_B 上,利用外准则进行中间待选模型的选留,但是如何快速选取误差较小(即可决系数较大)的模型是需要解决的问题。

为此有 2 种可选方案(快速选择法和大小堆法)来快速选择以实现外准则,使每次拟合的参数都是上次拟合的前 k 个最佳模型(即选择前 k 个最大可决系数的对应拟合方程,或前 k 个拟合误差最小的拟合方程)。这 2 种算法均可用来选择前 K 个最大(小)数。

4.1 快速选择算法

实际上快速选择算法跟快速排序算法的思想几乎一样。其基本思想(以查找前 k 个最大数为例)是:从 n 个数中,随机选取一个数 x(随机选取枢纽元,可以做到线性期望时间 $O(n)$的复杂度),将数组划分为 s_a 和 s_b 两部分,使得 $s_a \leqslant x \leqslant s_b$。如果要查找的 k 个元素小于 s_b 的元素个数,则返回到 s_b 中较大的 k 个元素,否则返回 s_b 中所有的元素 s_a 中大的 $k-|s_b|$ 个元素。同上述过程一样,运用类似快速排序的划分的快速选择算法寻找最大的前 k 个元素,这样平均复杂度为 $O(n)$,最坏复杂度为 $O(n^2)$。

不过值得一提的是,这个快速选择算法是选取数组中"中位数"作为枢纽元,而不是随机选取枢纽元。

4.2 大小堆方法

大小堆法的基本思路是:先用最大堆初始化数据集,将堆顶元素与堆中最后一个元素互换(即堆顶元素下移),并将原来的堆顶元素取出放到有限队列中,再用最大堆初始化剩下的数据,将堆顶元素与堆中最后一个元素互换,并将原来的堆顶元素取出放到有限队列中。依此类推,将堆顶元素下移 k 次后,优先队列中储存着原数据集中前 k 个最大的元素。然后从队列中取出这 k 个元素,这种算法的复杂度为 $O(n \lg k)$。

4.3 两种算法比较

在用 GMDH 处理数据时,往往处理相当多的数据,所以用 10 000 000 数据去选择前 K 个最大数,如图 1,其中 quick sort 是采用快速选择法,max heap 是采用最大堆算法。

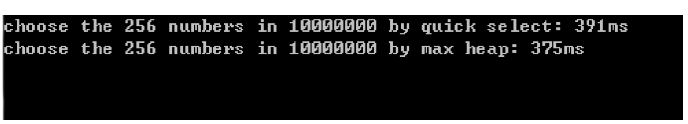

图 1 快速选择算法和大小堆法比较

4.4 实验结果

将上述改进加入到 GMDH 拟合算法中,在不改变原来的准确度的情况下,很大程度上降低了 GMDH 算法的时间复杂度。具体情况对比如图 2 所示。

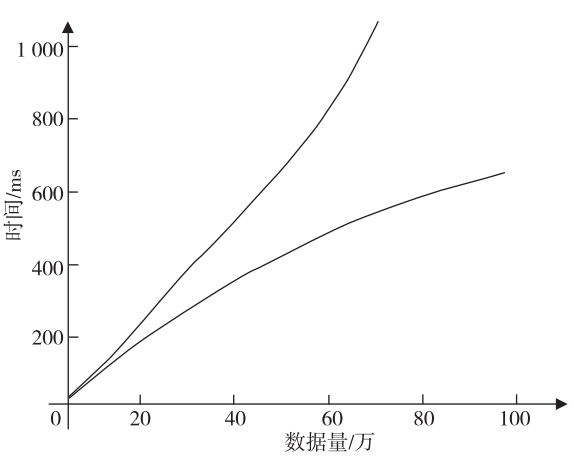

图 2 GMDH 算法的时间复杂度比较

5 结论

本文就 GMDH 算法在计算过程中,对计算量较大的两点做了改进,使得它在处理海量数据时,仍能较为快速准确地给出最佳模型,并作出最合理的短期预测。经过改进后,GMDH 算法的时间复杂度有明显的改善,使得 GMDH 算法在实际应用中性能增强。

参考文献:

[1] Ivakhnenko A G.Heuristic self-organizing in problems of engineering cybernetics[J]. Automatic,1967,6:207-219.

[2] Creswell J W,Hanson W E,Plano V L C,et al. Qualitative research designs: selection and implementation [J].The Counseling Psychologist,2007,35(2):236-264.

[3] Manuel B,Floyd R W,Pratt V R,et al. Time bounds for selection[J]Journal of Computer and System Sciences,1973,7(4):448-461.

[4] 张宾,贺昌政. GMDH 算法的终止法则研究[J]. 吉林大学学报:信息科学版,2005(3):257-262.

Zhang B, He C Z. Research on stopping criterion of GMDH [J]. Journal of Jilin University: Information Science Edition, 2005(3): 257-262.

[5] Zaychenko Y P, Kebkal A G, Krachkovskii V F. The fuzzy group method of data handling and its application to the problems of the macroeconomic indexes forecasting [J]. Systems Analysis Modelling Simulation, 2003, 13 (10): 1321-1329.

[6] 鲁茂,贺昌政,李慧. 线性 GMDH 参数模型的无偏估计研究[J]. 统计研究,2009(6):92-97.
Lu M, He C Z, Li H. The linear unbiasedness property of linear GMDH parametric model [J]. Statistical Research, 2009(6): 92-97.

[7] 鲁茂. 改进的 GMDH 算法及其应用[J]. 软科学,2008(4): 17-20.
Lu M. Improved GMDH algorithm and its application [J]. Soft Science, 2008(4): 17-20.

[8] 张秋菊,朱帮助. 一种改进的 GMDH 算法[J]. 郑州大学学报:理学版,2010(1):9-13.
Zhang Q J, Zhu B Z. A kind of improved GMDH algorithm [J]. Journal of Zhengzhou University: Natural Science Edition, 2010(1): 9-13.

[9] 李庆扬,王能超,易大义. 数值分析[M]. 北京:清华大学出版社,2009.
Li Q Y, Wang N C, Yi D Y. Numerical Analysis [M]. Beijing: Qinghua University Press, 2009.

[10] 鲜思东,潘显兵,胡学刚,等. 概率论与数理统计[M]. 北京:科学出版社,2010.
Xian S D, Pan X B, Hu X G, et al. Probability theory and mathematical statistics [M]. Beijing: Science Press, 2010.

[11] 冯兰刚,赵国杰,焦彦臣. 基于改进 GMDH 和 AC 聚类算法的竞争力模型研究——钢铁行业竞争力分析[J]. 西安电子科技大学学报:社会科学版,2011,21(2):1-5.
Feng L G, Zhao G J, Dai Y C. The research competit-iveness model and improved GMDH clustering algor-ithm based on AC [J]. Journal of Xidian University: Social Science Edition, 2011, 21(2): 1-5.

Research on the Method of Rapid Processing Data Packets

WANG Aihua

(Department of Humanity & Culture Art, Nanchong Professional Technical College, Nanchong Sichuan 637000, China)

Abstract: In the proper termination fitting condition, GMDH algorithm is relatively weak in processing large data. From the nature of GMDH computing, we solve the efficiency of the algorithm, an objective with fast inverse matrix to achieve a fast solving equation gives the algorithm, and the use of extrapolation standard, reasonable selection of fitting end conditions and selecting the best model method. In does not change the accuracy of the GMDH algorithm situation, study on how to improve the computational efficiency of GMDH.

Key words: the fast data processing algorithm; least square method; fitting principle; residual; inverse matrix

(责任编辑 黄 颖)