文章编号: 1001-2486 (2000) 04-0051-06

基于算子空间的公式发现算法研究*

赵新昱, 陈文伟, 何义

(国防科技大学人文与管理学院, 湖南 长沙　410073)

摘　要: 在FDD 提出的人工智能技术与曲线拟合技术结合的公式发现系统的基础上, 提出了新的基于算子空间的公式发现算法, 并在算法研究的基础上设计实现基于算子空间的可视化公式发现系统, 该系统通过算子空间概念的引入, 简化了算子空间的规则, 同时引入导数规则、误差规则以及终止规则, 丰富了知识库内容。通过以上改进, 和 BACON 和 FDD 相比, 公式发现的形式更广, 复杂度更高。文章最后给出了应用实例以及公式发现的结果。

关健词: 算子空间; 公式发现; 规则; 曲线拟合
中图分类号: TP301　　文献标识码: A

Formula Discovery from Data Based on Operators Space

ZHAO Xin-yu, CHEN Wen-wei, HE Yi
(College of Humanism and management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Based on the research of AI involved in FDD and data approach, The paper has presented a new algorithm for formula discovery from a large amount of experimental data. And then we build up the system by using the algorithm based on space of operators. The basic rules of operators are simplified, and the other three rules: derivative rule, inaccuracy rule and end rule, are introduced, which extend the knowledge lib and make it easy to use the rules in lib. Through the above improvement more forms and complexity of functional formulas are found than BACON and FDD system. In the end, the simulation result is given.

Key words: operators space; formula discovery; rules; data approach

很多实际问题可以归结为如图 1 所示抽象系统[6]。若已知有限个输入输出对 $\{(x_i, y_i); i = 1, 2, \cdots, n\}$, 要求构造系统结构, 则是一个建模问题。应用到数据开采领域, 称为公式发现。

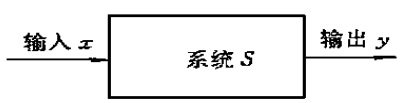

图 1　系统抽象描述
Fig. 1　System abstract description

公式发现的主要思想是, 利用科学实验中得出的大量测量数据, 去求得自变量与因变量的一个近似公式, 便于对实际问题进行分析和研究[1,3,4]。

其具体步骤是:

(1) 通过观察实验现象获得大量数据;

(2) 凭借领域知识对大量数据进行分析;

(3) 根据分析结果, 提出该数据的假设原型;

(4) 通过计算对原型进行验证, 利用每次验证结果, 提出新的假设原型, 直至得出理想结果。

公式发现中比较典型的方法有: 曲线拟合, 科学定律发现系统 BACON, 数学概念发现系统 AM 以及经验公式发现系统 FDD (Formula Discovery from Data) 等。曲线拟合主要应用代数多项式形式, 它的系数由最小二乘原理建立正规方程组求出, 通过引进正交多项式、勒让德多项式等, 数据拟合方法是能对一般实验数据找到满足精度的逼近公式。但代数多项式存在一定的问题, 首先容易出现“病态”; 其次问题描绘不直观; 随着代数多项式次数的升高, 问题越来越复杂。

BACON 系统是运用人工智能技术从实验数据中寻找其规律性比较成功的一个系统。它是 Pat Lan-

* 收稿日期: 1999-12-14
基金项目: 国家自然科学基金资助项目 (79670019)
作者简介: 赵新昱 (1973-), 男, 博士生。

gly 于 1980 年研制的。它运用数据驱动方法，整个学习程序由多个精炼算子组成，系统使用探索知识对提供的训练例进行分析，决定选用哪个精炼算子。BACON 系统的思想是，系统反复地考察数据并使用精炼算子创造新项，直到创造的这些项中有一个是常数时为止。于是一个概念就用“项= 常数”的形式表示出来，其中“项”为变量运算的组合而形成的表达式。

数学概念发现系统 AM 采用模型驱动方式，在多种搜索方法指导下在数据领域中搜索，从集合、表、项等 1000 多个基本数学概念出发，并使用具体化、一般化、类比、复合等操作，产生新的数学概念。

FDD 系统是将人工智能技术与数据拟合技术结合起来，通过启发式搜索，找到满足精度要求的经验公式。它是从大量数据中发现经验公式，运用人工智能以及曲线拟合技术，逐步完成任意函数的任意组合。在实际应用中有一定的效果。

本文在以上研究成果的基础上，提出了基于算子空间的公式发现算法，对几个具体应用实例进行分析，并进一步对发现的结果进行了验证。

1 算子空间的构造

算子空间由算子组成，算子不仅包括原型算子，组合算子和嵌套算子，同时引进导数算子，通过对算子的扩充，增强了发现公式的复杂度。算子空间的构造如下:

算子空间 Ω 的定义: $\Omega = < PO, V, C >$, $PO = \{f_1^{n1}, f_2^{n2}, \cdots f_m^{nm}\}$ 是一个有穷算子集,$f_i^{ni}(i = 1, \cdots, m)$ 是 ni 元算子。

算子集可以包括:

- 算术运算(如+ , - , * , / 等);
- 数学函数(如 1, x^1, x^2, $x^{1/3}$, sin, cos, exp, log 等);
- 导数算子(如差分算子、差商算子等), 设给出的测量数据为:

i	1	2	…	N
x	x_1	x_2	…	x_n
y	y_1	y_2	…	y_n

则, 一阶差分: $\Delta x_k = x_{k+1} - x_k$; $\Delta y_k = y_{k+1} - y_k$; $(k = 1, 2, \cdots, n-1)$

二阶差分: $\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k$; $\Delta^2 x_k = \Delta x_{k+1} - \Delta x_k (k = 1, 2, \cdots, n-2)$

一阶差商 $\delta y_k = (y_{k+1} - y_k)/(x_{k+1} - x_k)$ $(k = 1, 2, \cdots, n-1)$

二阶差商 $\delta^2 y_k = (\delta y_{k+1} - \delta y_k)/(x_{k+2} - x_k)$ $(k = 1, 2, \cdots, n-2)$;

- 其它与问题有关的函数, 如自定义函数等。

$V = \{v_1, v_2, \cdots, v_k\}$ 是一个有穷变元集。$C = \{c_1, c_2, \cdots, c_k\}$ 是一个有穷常元集。

记 $E = V \cup C$, 并满足条件:

⑴ 若 $e \in E$, 则 $e \in \Omega$;

⑵若 $f^{(n)} \in PO$, $e_i \in E$, $i = 1, 2, \cdots, n$, 则 $f^{(n)}(e_1, e_2, \cdots, e_n) \in \Omega$;

⑶若 $p_1, p_2, \cdots, p_n \in \Omega$, 则对于 $f^{(n)} \in PO$, $f^{(n)}(p_1, p_2, \cdots, p_n) \in \Omega$。

满足以上条件的算子空间是封闭的，所以，对于在算子空间上的任意算子组合，仍然在算子空间中，这样为应用计算机对算子空间的处理提供了可以迭代的前提。

2 公式生成流程

传统的公式生成搜索树，按照对变量 x 和变量 y 分别应用算法空间中的算法进行, 每应用一次时, 就进行误差计算, 选择误差最小的进入下一次迭代。在应用到具体问题时, 应用这种方法不能保证上一次好的公式能进入下一次迭代, 导致结果出现偏差, 不能找到好的公式, 所以要对公式生成算法进行改

进。在生成搜索树时, 同时对变量 x 和变量 y 应用算法, 由于算法空间包含算法 $y = x$, 所以改进的搜索树包含原有的公式生成算法。改进的公式生成算法的搜索树如图 2 所示。

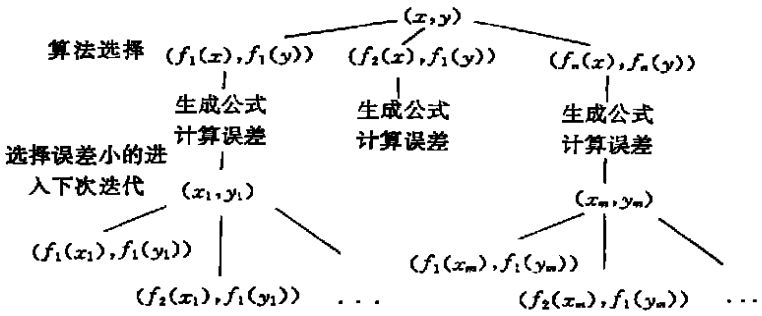

图 2 公式生成算法搜索树

Fig. 2 Search tree of formula discovery method

在公式生成中引入曲线拟合方法, 即求线性数学模型:

$$f(y) = b_1\varphi_1(x) + b_2\varphi_2(x) + \cdots + b_n\varphi_n(x) \tag{1}$$

其中函数 f 和 $\Phi_1, \cdots, \Phi_n$ 为算法空间中选择的函数, 现取 m 个实验数据点 (x_i, y_i), $i = 1, 2, \cdots, m$, 把 m 个点代入方程(1) 中, 得到一矛盾方程组, 要求 $m > n$, 应用最小二乘法确定各个参数 $b_1, b_2, \cdots, b_n$, 使各点的误差平方和最小:

$$Q = (Y - y)^2 = (Y - f^{-1}(b_1\varphi_1(x) + b_2\varphi_2(x) + \cdots + b_n\varphi_n(x)))^2 = \min \tag{2}$$

由数学分析求极值的方法, 求出 Q 对 b_k 的偏导, 并令其等于 0, 得

$$\frac{\partial Q}{\partial b_k} = \sum_{i=1}^{m} 2(y_i - f^{-1}(\sum_{j=1}^{n}\varphi_j(x_i)))\left(\frac{\partial(f^{-1}(\sum_{j=1}^{n}\varphi_j(x_i)))}{\partial b_k}\right) = 0 \quad (k = 1, 2, \cdots, n)$$

求这组方程得解 $\{b_k\}$, 即可求出拟合公式。

引入导数算法后, 生成的公式类似于: $y' = f(x)$, 经过处理得:

$$y = \int f(x)\,\mathrm{d}x + C \tag{3}$$

应用最小二乘法求 C 值。要求:

$$Q = (Y - y)^2 = (Y - \int f(x)\,\mathrm{d}x + C)^2 = \min$$

Y 是实际观测数据值, 为使 Q 最小, 求 Q 对 C 的偏导, 同样令其等于 0 , 得:

$$\frac{\partial Q}{\partial C} = \partial(\sum_{i=1}^{m}(y_i - \int f(x)\,\mathrm{d}x + C)^2)/\partial C = 0 \quad (k = 1, 2, \cdots, n)$$

经计算得:

$$C = \frac{1}{m}\sum_{i=1}^{M}(y_i - y_i^*) \tag{4}$$

即得生成公式。

3 基本规则

系统中的知识采用产生式规则表示形式 (if …then …), 由于定义了算子空间, 所以对算子规则形式可以简化为:

1) 算子嵌套规则

对某一变量 x_i, 取算子空间中的一算子后, 与另一变量 x_j 进行线性组合, 若为常数, 则建立关系式:

$$c_1 f_j^{\,i}(x_i) + c_2 f_{j'}^{\,i'}(x_j) = c_3 \qquad f_j^{\,i}, f_{j'}^{\,i'} \in PO,\ c_1, c_2, c_3 \in C \qquad \text{(规则 1)}$$

对规则 1 嵌套或递归使用，将形成变量的任意组合。

2) 导数算子规则

差分判定规则（设 $\Delta x_i (i = 1, 2, \cdots n-1)$ 为定值）：

若 $\Delta y_i =$ 定值，则方程为 $y = a + bx$；

若 $\Delta^2 y_i =$ 定值，则方程为 $y = a + bx + cx^2$；

差商判定规则：

若 $\Delta y_i / \Delta x_i =$ 定值，则方程为 $y = ax + b$；

若 $\Delta \log(y_i) / \Delta \log(x_i) =$ 定值，则方程为 $y = ax^b$；

若 $\Delta \log(y_i) / \Delta x_i =$ 定值，则方程为 $y = ab^x$；

若 $\Delta(x_i y_i) / \Delta x_i =$ 定值，则方程为 $y = a + b/x$；

若 $\Delta(x_i / y_i) / \Delta x_i =$ 定值，则方程为 $y = x/(ax + b)$；

若 $\Delta y_i / \Delta(x_i)^2 =$ 定值，则方程为 $y = a + bx^2$；

3) 误差规则

- 误差最小规则：选择误差最小的公式进入下一次迭代；
- 误差收敛规则：保留误差减小的搜索方向，上一次迭代的误差大于目前的误差，则对于这一搜索方向予以保留。

系统主要通过误差规则来确保算法的收敛性，即下一次迭代的误差一定小于上一次迭代的误差。

4) 终止规则

终止准则由两部分组成，一是强制终止，另一个是自然终止，强制终止通过对算法参数的设定，主要是通过对迭代次数的设定完成终止准则；自然终止有两种情况组成，一种是找到一组满足给定误差的公式，另一种情况是判断出误差不收敛，即采用该方法不适合求解给定的问题。

4 基于算子空间的公式发现系统体系结构

系统体系结构如图 3 所示。

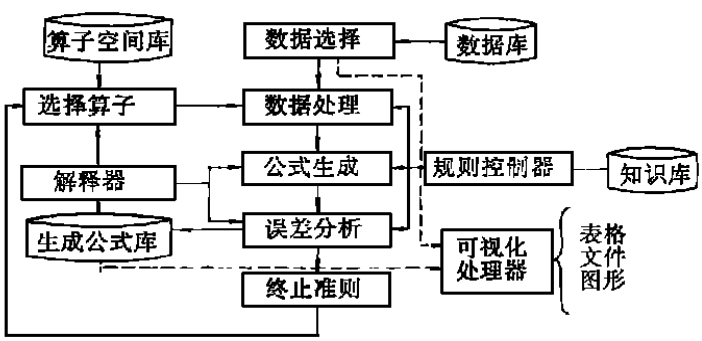

图 3 基于算子空间的系统体系结构图

Fig. 3 The formula discovery system architecture based on operators space

1) 数据选择：从数据库中选择数据；

2) 选择算子：从算子空间库中选择算子；

3) 数据处理：应用算子对数据进行处理；通过规则生成器控制完成公式基本形式，数据处理还包括数据预处理，对数据进行去噪等操作；

4) 公式生成：应用数学拟合方法，通过规则生成器的控制，完成公式参数的求解和误差的求解；

5) 误差分析：应用最小二乘法求得参数值后，就可以进行误差分析。通过对误差的分析，比较各个公式的优劣，对公式进行选优，好的公式输入生成公式库中保存。另外误差分析的另一项工作是，通过对每次迭代的误差走势进行分析，判断问题是否适合求解。

6) 解释器: 解释机制完成对算法空间元素的统一解释, 解释器不仅包括算法基本原型的解释, 还包括对数据的求导, 对函数的求反, 公式求积分, 对复合函数的解释, 对生成公式的运算等。

7) 终止准则: 通过手工设定和系统缺省判定规则, 控制迭代次数以及生成公式的误差;

8) 规则控制器: 从知识库中读取知识, 并按照知识库的规则, 完成对数据处理, 公式生成和误差分析的控制;

9) 可视化处理: 可视化处理包括图形输出和表格输出, 系统生成若干公式后, 通过此模块对生成公式进行比较和检验, 这样用户可以通过形象化的输出得出问题的描述特征。

系统共有四个库: 数据库、算子空间库、生成公式库和知识库。数据库由用户建立; 算子空间库存储算子信息; 生成公式库对系统产生的合乎要求的公式进行存储; 知识库存储规则。

5 应用实例

1) 利用开普勒的近似数据进行发现, 共发现三个公式:

$$y^3 = 0.00 + 1.00x^2 \quad \ln(y) = 0.00 + 1.535\lg(x) \quad \lg(y) = 0.00 + 0.667\lg(x)$$

在文献 [3] 中的 FDD 系统中发现一个公式 $\lg(y) = 0.00 + 0.667\lg(x)$。

2) 组合算子公式的发现: X, Y 为样本数据, y 为发现的公式计算值。

X	1.01	2.07	2.98	3.89	5.02	6.03	6.98	8.01	9.04	9.99	11.02	12.01	12.97
Y	4.61	10.51	14.65	14.61	11.08	10.2	12.6	18.27	23.3	24.46	22.08	19.72	20.93
y	4.667	10.662	14.248	14.524	11.741	10.383	12.679	18.263	23.174	24.257	22.045	19.965	21.115

利用导数算子规则发现公式如下:

$$y' = 1.5160419854043x + 4.342026905719\sin(x)$$

消除导数关系时利用公式 (4) 求常数 C, 最后的公式为:

$$y = 1.5160419854043x - 4.3420264905719\cos(x) + 5.4447611299494 \quad \text{误差: } 0.04829553$$

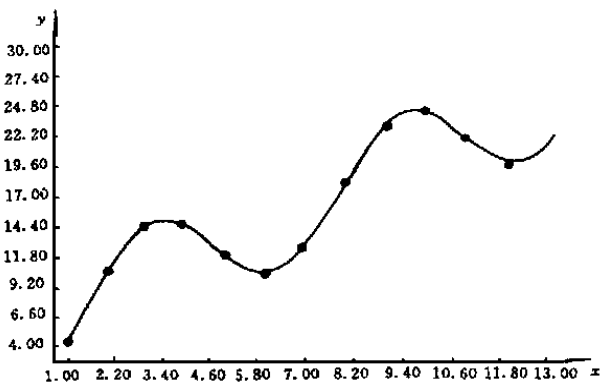

图 4 组合算子公式的发现

Fig. 4 Discovery of combined operators formula

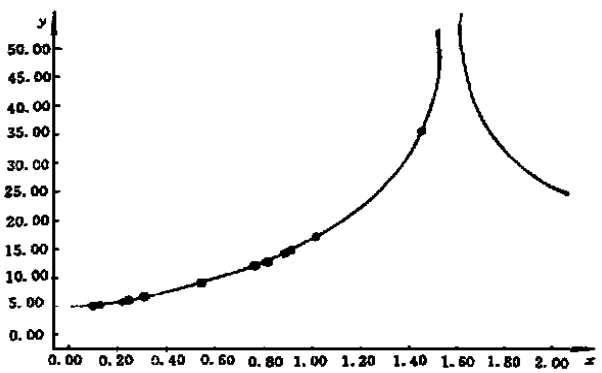

图 5 复合算子公式的发现

Fig. 5 Discovery of compound operators formula

3) 复合算子公式的发现, 数据如下:

X	0.10	0.12	0.23	0.25	0.30	0.26	0.55	0.76	0.81	0.89	0.91	1.01	1.44	1.50
Y	5.146	5.288	6.156	6.329	6.782	6.417	9.532	12.588	13.443	14.936	15.337	17.53	35.81	43.02
Y_1	4.899	5.044	5.924	6.10	6.561	6.190	9.371	12.51	13.385	14.921	15.334	17.59	36.50	43.98
Y_2	6.65	6.66	6.67	6.80	6.92	6.82	8.38	11.236	12.204	14.021	14.530	17.43	37.91	41.98
Y_3	5.185	5.310	6.07	6.228	6.636	6.306	9.268	12.525	13.491	15.223	15.696	18.33	37.0	40.99

发现公式 1: $Y_1 = 5.941028061x + -11.63530118\log(|\cos(x)|) + 4.247142360$, 如图 5 所示。

该公式的误差为: 0.09509156

另外还发现两个公式:

公式 2: $Y_2 = 6.639005246 + 10.47187751x^3$, 误差为 0.29077。

公式 3: $\mathrm{sqrt}(Y_3) = 0.92690791 + 1.221648810e^x$, 误差为 0.18251。

6 结束语

基于算子空间的公式发现算法不论从发现公式的复杂度, 公式形式以及公式的数量上都比 BACON 和 FDD 系统有明显的进步, 同时简化算子规则, 增加误差规则, 导数规则和终止规则, 有利于算法的实现。从发现公式的误差来看, 其也要比 FDD 要好。另外改进了搜索树的搜索方法, 避免了盲目剪枝。

算子空间的范围和算子的定义根据具体问题的不同而不同, 如在二进制上定义算子空间, 则可能是 {and, or, xor}。另外一个问题可以应用不同的算子空间进行公式发现。

实例 3 共发现三个公式, 其中两个公式有相同的发展趋势, 另外一个公式从表现图上来看, 有和其它两个不同的特性, 这正说明了应用该方法的有效性, 能够发现不同类型的公式。

参考文献:

[1] Langley P W. BACON: A Production System that discovers Empirical Laws [J], IJCAI 1977.
[2] 王福明等. 应用数值计算方法 [M]. 北京: 科学出版社, 1992.
[3] 陈文伟. 智能决策技术 [M]. 北京: 电子工业出版社, 1998.
[4] Michalski R S 等. 机器学习- 实现人工智能的途径 [M]. 北京: 科学出版社, 1992.
[5] 陈文伟等. 可视化机器发现的研究 [J]. 国防科技大学学报, 增刊, 1995.
[6] 张维明等. 信息系统建模技术 [M]. 北京: 电子工业出版社, 1998.

(上接第 40 页)

综上所述, 通过进一步提高离子交换膜的交换容量, 改善复合膜的工艺条件, 有望进一步提高膜分离体系的效率。实验的三种膜分离情况的汇总如图 5 所示。可以看出本方案在提高原分离膜的分离效率方面有较明显的作用, 即当膜的分离效率较低时, 可提高分离通量; 在膜的分离效率较高时, 可大幅度降低膜分离体系的分离膜面积。

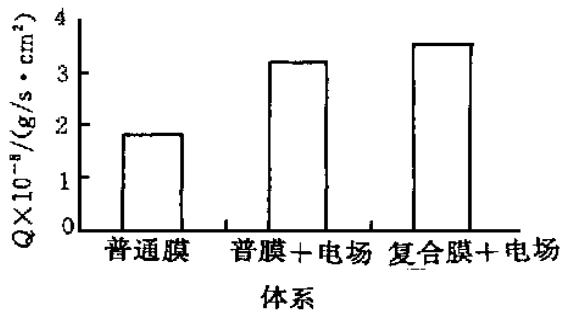

图 5 三种分离方式的改进效果比较

Fig. 5 Comparison of three separation methods

3 结论

经过上述研究可以得到以下结论:

(1) 利用离子交换膜是分离低分压 CO_2 混合气体中 CO_2 气体的一种有效的方法。

(2) 离子交换膜的处理工艺条件对其 CO_2 的分离性能有很大的影响。

(3) 含有亲水物质的离子交换复合膜可提高膜的 CO_2 通量。

(4) 低电压电场驱动可较大幅度地提高离子交换膜的 CO_2 通量。

参考文献:

[1] Vorob' ev A V, Khim T O. Recent advance in CO_2 separation [J]. Technol., 1985; 19 (4): 544.
[2] Ward W J. ㈣Recent Developments in Separation Science [M]. Vol. 1 1972, N. N. Li ED., CRC Press, Cleveland.
[3] Way J D, Noble R D. Facilitated transport of CO_2 in ion exchange membranes [J]. ALChE Journal, 1987, 33 (3): 480-484.