

◎数据库、信号与信息处理◎

商空间粒度预测方程

王兴野, 韩 燮

WANG Xing-ye, HAN Xie

中北大学 电子与计算机科学技术学院, 太原 030051

School of Electronics and Computer Science and Technology, North University of China, Taiyuan 030051, China

E-mail: s20050321@mail.nuc.edu.cn

WANG Xing-ye, HAN Xie. Forecast function of quotient space granulation. Computer Engineering and Applications, 2008, 44(17): 125-127.

Abstract: Based on quotient space granulation theory, combine many affects, such as granular computing, the increasing granular information and information abstracting of granulation. Moreover, rebuild the topology relations of the abstracting information based on the topology of yield granulation. Hence the increasing information accord with a topology. Finally, point out the quotient space theory forecast function based on gray theory and its methods. This is a new practice of information forecast.

Key words: quotient space theory; forecast function; granular computing; granular information; information abstracting; gray theory

摘 要: 以商空间粒度计算理论为基础, 综合考察了粒度计算, 粒度上信息的增长, 以及粒度集合上信息的白化问题。进一步在原粒度拓扑关系基础上建立了粒度白化信息的拓扑结构, 使信息增长符合一定的拓扑关系。最后结合灰色理论的方法论建立了商空间应用领域的预测方程, 是信息预测的新尝试。

关键词: 商空间理论; 预测方程; 粒度计算; 粒度信息; 信息白化; 灰色理论

DOI: 10.3778/j.issn.1002-8331.2008.17.037 **文章编号:** 1002-8331(2008)17-0125-03 **文献标识码:** A **中图分类号:** TP301.4

1 引言

预测就是以系统的过去为依据, 并参照当前出现的各种情况去推测、了解事物的未来。要想做出正确的预测, 就必须对客观事物的过去和现状进行深入研究和科学分析。事物的发展是连续的, 信息的积累是逐步的。过去、现在和未来之间必然有着某种联系, 预测者要想找出这种规律, 预见未来的发展趋势, 就必须使用一种逻辑结构, 采用一定的信息技术或人工智能的手段, 以达到预测的目的。一般的预测方法或者受主观因素影响较大。或者对数据量的要求较大, 而且希望数据具有典型分布。这在实际上很难实现, 以致预测精度不高, 有时甚至出现反常现象, 而灰色预测方法则可不受这些限制。灰色预测理论是灰色系统理论的重要组成部分, 是基于灰色理论的一种预测方法。它用生成模块建立微分方程模型, 可从少量的、离散的、杂乱无章的数据中找出规律性。并且具有良好的时效性。结合商空间信息结构性的优点和灰色理论数据少, 离散, 杂乱无章的特点构造新的带信息结构的预测方程。

2 商空间粒度理论

商空间粒度理论是关于复杂问题求解的空间关系理论, 其主要内容包括复杂问题的商空间描述、商空间的粒度计算、粒

度的合成和粒度空间关系的推理等。一般用三元组 (X, f, T) 描述一个复杂问题。 X 表示问题的论域; $f(x)$ 表示论域中元素的属性, 属性函数 $f(x)$ 可以是单值也可以是多值的; T 表示论域的结构, 即论域 X 中各元素的相互关系。

粒度概念可以用子集来表示, 粒度的大小体现为不同的子集, 构成空间的一个划分, 即商空间。粒度集合 $([X], [f], [T])$ 表示 (X, f, T) 的一个划分。不同的粒度划分就构成不同的商空间。粒度计算就是研究给定商空间上的各种子集合之间的关系和转换, 以及对同一问题取不同的粒度, 并在对不同粒度的研究中, 综合获取原问题的解。把对问题 (X, f, T) 的求解转化为求不同的粒度 $([X], [f], [T])$ 上的解。

商空间粒度划分是提取信息的方法。对于不同的划分, 粒度大小不同, 所携带的信息多少也不相同。对于细粒度的划分粒度上具有的信息相对较确定, 粗粒度上所携带的信息模糊, 但相对较丰富。如 R_1 是 $([x]_1, [f]_1, [T]_1)$ 上的关系, R_2 是 $([x]_2, [f]_2, [T]_2)$ 上的关系, 当 $R_1 > R_2$ 时, 有 $([x]_1, [f]_1, [T]_1) \supset ([x]_2, [f]_2, [T]_2)$ 。在由 R_2 向 R_1 转化的过程中, 信息在逐渐增加、膨胀。商空间的半序结构是粒度合成与分解的结果, 也是信息合并分离的结果。信息在商空间下构成半序集。由细粒度到粗粒度信息在增长, 同时也在逐渐的变的不确定。信息模糊化是系统发展

作者简介: 王兴野(1980-), 男, 硕士研究生, 主研领域: 商空间粒度计算、信息预测; 韩燮(1964-), 女, 博士, 教授, 主研领域: 数据挖掘、粒度计算。

收稿日期: 2007-08-23 **修回日期:** 2007-12-06

的趋势。

在商空间理论中,系统信息模糊化不是突变的,它伴随着商空间粒度的转化发生。提取信息增长的拓扑关系,对于系统发展趋势的了解是有益的。下面探讨应用构造树提取这一信息。

3 商空间结构树构造算法

根据商空间理论的思想,一个概念可以在论域的某个商空间(某个层次)上进行描述,使用 Gaussian 型函数定义商空间距离的模糊聚类 FCluster 算法。其算法思想是:用样本间属性的距离表示信息粒度进行聚类,所得到的不同粒度的商空间即是不同的聚类结果,并可以从不同角度、不同层次观察聚类的结果,即以不同粒度来描述样本集,在不同层次反映系统信息。

设给定 m 个样本的样本集为 $X=\{x_i, i=1,2,\dots,m\}$, $X \in R^n$ 是 n 维欧氏空间的点集。设距离函数 $d(x,y)=1-\exp(-\|x-y\|^2/2\sigma^2)$, $x,y \in X$, σ 表示 RBF 函数的宽度参数。这样的距离函数具有以下一些性质:

(1) $\forall x,y \in X$ 有 $d(x,y) \in [0,1] \in [0,1]$, 是一个严格单调增加函数;

(2) $\forall x,y \in X$, 当

$$d(\|x-y\| \rightarrow \infty, d(x,y) \rightarrow 1$$

(3) 正定性: $d(x,y) \geq 0, d(x,y)=0$ 当且仅当 $x=y$;

(4) 对称性:

$$d(x,y)=1-\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)=1-\exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right)=d(y,x)$$

依据商空间模糊粒度计算理论,把 Gaussian 型函数定义为商空间的距离函数,对距离 d 小于粒度 d_{λ} 的样本聚为一类,从而设计出模糊聚类 FCluster 算法。

应用 FCluster 算法进行商空间层次树 S 的构造。选择一系列 d_{λ_i} 使 $d_{\lambda_1} > d_{\lambda_2} > \dots > d_{\lambda_l}$, 对于 $d_{\lambda_i} (i=1,2,\dots,l)$ 的选取依据经验和专家建议选择。对数据集 X 进行 l 次聚类运算,由 d_{λ_i} 诱导出等价关系 R_{λ_i} , 通常等价关系 $R_{\lambda_i} \supseteq R_{\lambda_{i+1}}$ 并不一定成立,及 $([x]_{\lambda_i} \cap [y]_{\lambda_i} \cap [z]_{\lambda_i})$ 中的元素与 $([x]_{\lambda_{i+1}} \cap [y]_{\lambda_{i+1}} \cap [z]_{\lambda_{i+1}})$ 中的元素不都存在包含关系。应用商空间构造半序集的方法调整 R_{λ_i} 的结构得到 R'_{λ_i} , 使得 $R'_{\lambda_1} > R'_{\lambda_2} > \dots > R'_{\lambda_l}$, 由此问题 X 的完整半序集构造成功。若取 $R'_{\lambda_l}=1$, 则由 $R'_{\lambda_1} > R'_{\lambda_2} > \dots > R'_{\lambda_l}$ 半序结构构造得到树结构 S 。

4 灰色预测模型

灰预测模型原理主要是对原始数列进行处理,一般进行一次累加处理,其目的主要是为了弱化原始数据的波动性,可以证明经过累加处理后所生成的序列能够表现出明显的指数规律,这时可以用一生成序列的后续值,最后再累减还原即得到原序列的预测值。其预测过程如下:

(1) 给定原始数据序列

$$x^{(0)}(t)=\{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$$

(2) 对 $x^{(0)}(t)$ 作一次累加,求得一次累加生成数列

$$x^{(1)}(k)=\{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n)\}$$

(3) 构造数据矩阵 $B, y_n^{90^\circ}$

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(1)+x^{(1)}(2)) & 1 \\ -\frac{1}{2}(x^{(1)}(2)+x^{(1)}(3)) & 1 \\ \dots & \dots \\ -\frac{1}{2}(x^{(1)}(n-1)+x^{(1)}(n)) & 1 \end{bmatrix}$$

$$y_n = [x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(n)]^T$$

(4) 求参数向量 \hat{a}

$$\hat{a} = [a, u]^T = (B^T B)^{-1} B^T y_n$$

(5) 将参数向量 \hat{a} 代入时间响应函数

$$\hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a}$$

(6) 确定还原模型

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k)$$

系统的预测值可由上式得到。

5 商空间粒度灰色方程

5.1 商空间粒度信息白化及拓扑信息树建立

商空间粒度灰色方程的构造与灰色方程的构造有相似之处也有不同点。灰色方程构造的基础是灰色数列,灰色数列由灰数白化得到,灰数的白化方法很多,如平均值法、权重法,都是对灰信息的白化处理。在灰色理论中,信息的携带者是经过信息白化的数列,而在商空间理论中,信息的携带者是以粒度集合为单位的。灰色数的白化是在不确定中寻找确定,与在模糊集上找到能代表整个集合信息的元素相似。因此商空间粒度集合上的信息白化是可行和必要的。粒度信息的白化信息将是基本的信息单位。与灰色模型的数据有顺序拓扑关系相同,商空间粒度集合上的拓扑关系较复杂,找到粒度白化信息上的拓扑关系将是模型建立的重点。商空间粒度划分的目的是:从不同层次,不同侧面研究问题。从不同粒度上研究问题,不同的划分可以构成粒度上的半序结构。半序结构反映信息之间的关系,在这种关系上构建系统模型,符合信息增长的规律。商空间粒度灰色方程的拓扑关系就是建立在半序结构上的。由上面商空间结构树构造算法构造的拓扑树 S , 构成了 X 的一个信息增长拓扑关系。

商空间粒度信息的白化与商空间数列拓扑关系的建立是同时进行的。对商集 $[X]_i = \{[x_0]_i, [x_1]_i, \dots, [x_k]_i\}$ 的子集合求均值得到 $[X]_i$ 的白化信数为数集 X_i 。其中 $[X]_i$ 中的元素为树 S 的叶子节点。用 X_i 中的元素代替相应的叶子节点,上层的节点为下层节点的和,由此构造出新的结构树 S_g, S_g 为有拓扑关系的白化信息树。其中根节点的值为 X_i 元素的和。从第 i 个叶子节点回溯到根节点,构成数列 L_i , 因有 k 个叶子节点可构造 k 个数列。数列 L_i 的数据由小到大排列。

5.2 建立灰色方程

分析结构树 S_g 以及数列 L_i , 与灰色模型不同的是商空间粒度信息白化数列 L_i 的增长步长为 $1, 2, 4, 8, 16 \dots$ 。设 L_i 中的数据为 $(x_i^{(0)}(0), x_i^{(0)}(1), x_i^{(0)}(2) \dots, x_i^{(0)}(m))$, 则 $(x_i^{(1)}(0), x_i^{(1)}(1),$

$x_i^{(1)}(2), \dots, x_i^{(1)}(m)$ 为 L_i 中的数据 t 次累加的数列。在此把数据 L_i 当作一次累加。

建立灰色方程; 设 $\frac{dx}{dt} = \frac{x_i^{(0)}(j) - x_i^{(0)}(j-1)}{2^{(j-1)} - 2^{(j-2)}}$, 得: $\frac{dx}{dt} = \frac{x_i(j)}{2^{(j-2)}}$;

灰色方程为: $\frac{x_i(j)}{2^{(j-2)}} + ax_i^{(0)}(j) = b$; $\frac{x_i(1)}{2^{-1}} + ax_i^{(0)}(1) = b$, $\frac{x_i(2)}{2^0} +$

$a_i x_i^{(0)}(2) = b, \dots, \frac{x_i(k)}{2^{(k-2)}} + ax_i^{(0)}(k) = b$ 。

写成矩阵:

$$Y = \Phi \theta$$

其中

$$Y = \left(\frac{x_i(1)}{2^{-1}}, \frac{x_i(2)}{2^0}, \dots, \frac{x_i(k)}{2^{k-1}} \right)^T$$

$$\Phi = \begin{pmatrix} -x_i^{(0)}(2) & 1 \\ -x_i^{(0)}(3) & 1 \\ \dots & 1 \\ -x_i^{(0)}(k) & 1 \end{pmatrix}, \theta = \begin{pmatrix} a \\ b \end{pmatrix}$$

用最小二乘法求解 θ 向量: $\theta = [a, b]^T = (\Phi^T \Phi)^{-1} \Phi^T Y$ 。

设 θ 的估计值为 $\hat{\theta}$, 可得误差序列: $E = Y - \Phi \hat{\theta}$, 其中 $\hat{\theta} = (\hat{a} \hat{b})^T$ 。

由此得到灰色方程 $\frac{dx_i^{(0)}(k)}{dt} + \hat{a} x_i^{(0)}(k) = \hat{b}$ 的离散形式解

$\hat{x}_i^{(0)}(k) = (x_i(k) - \frac{\hat{b}}{\hat{a}}) \exp(k-1) + \frac{\hat{b}}{\hat{a}} (k=0, 1, 2, \dots, m)$, $\hat{x}_i(k+1) =$

$\hat{x}_i^{(0)}(k+1) - \hat{x}_i^{(0)}(k) (k=0, 1, 2, \dots, m-1)$ 得到预测序列为 $\hat{x}_i^{(0)} =$

$[\hat{x}_i^{(0)}(1), \hat{x}_i^{(0)}(2), \dots, \hat{x}_i^{(0)}(m)]$ 。

树结构 S_g 有 k 个叶子节点, 因此可以建立 k 个灰色方程,

得到 k 个预测序列 $\hat{x}_i^{(0)} (i=1, 2, \dots, k)$ 。对上述 k 个预测值的综合是必要的, 平均值或权重均值法都是有效的方法。平均值:

$$\bar{x} = \left(\frac{\sum_{i=1}^k \hat{x}_i^{(0)}(1)}{k}, \frac{\sum_{i=1}^k \hat{x}_i^{(0)}(2)}{k}, \dots, \frac{\sum_{i=1}^k \hat{x}_i^{(0)}(m)}{k} \right)$$

权重均值:

$$\bar{x} = \left(\sum_{i=1}^k \frac{||x_i||}{|X|} \hat{x}_i^{(0)}(1), \sum_{i=1}^k \frac{||x_i||}{|X|} \hat{x}_i^{(0)}(2), \dots, \sum_{i=1}^k \frac{||x_i||}{|X|} \hat{x}_i^{(0)}(m) \right)$$

6 应用和分析

本文将商空间结构树构造算法应用于我国 1979 年到 2000 年的粮食产量数据, 数据集包含 22 个样本。σ 的选取与数据的样本值有关, 由于 $\|x-y\|^2$ 的值较大, 故选 $\sigma=1000$, 首先建立商空间结构树 S , 选择 $d_{A5}=0.2, d_{A4}=0.4, d_{A3}=0.6, d_{A2}=0.8, d_{A1}=1$, 在结构树的建立过程中, 最前一项 d_{A1} 一般取 1, 以得到根节点。当 $d_{A5}=0.2$ 由 FCluster 算法得到商集如表 1。

通常情况下, 由 d_{A1} 诱导出的 R_{A1} 关系在结构数的建立过程中不做调整。上面 $d_{A5}=0.2$ 划分了 8 个商集, 因此构造树有 8 个叶子节点, 可以建立 8 个商空间粒度灰色方程对系统 X 进行预测。继续应用 $d_{A4}=0.4, d_{A3}=0.6, d_{A2}=0.8, d_{A1}=1$ 划分商空间, 应

表 1 FCluster 算法得到的商集

聚类类别	聚类样本/万吨
聚类 1	33 211.5, 32 055.5, 32 502.0
聚类 2	35 450.0(孤立点)
聚类 3	38 727.5, 37 911.0, 39 151.0
聚类 4	40 731.0, 39 408.0, 40 298.0, 40 755.0
聚类 5	44 624.0, 43 529.0, 44 265.8, 44 510.0
聚类 6	45 648.8(孤立点)
聚类 7	46 217.5, 46 662.0
聚类 8	49 417.1, 50 453.5, 51 229.5, 50 838.6

注: 数据资料来源于《中国统计年鉴-2006》

用商空间半序集构造理论构造树型结构 S , 取 $d_{A1}=0.2$ 时的粒度信息白化值代替树 S 的叶子节点得到白化信息树 S_g , 树 S_g 最终应用于信息预测。

预测结果(表 2)表明模型能够预测较长的时期, 因为预测方程是建立在步长为 $2^n (n=0, 1, 2, \dots, l)$ 倍数逐渐增大基础上的, 这有利于保持过去和现在的信息, 减少模型求解过程中对原有信息的遗忘, 保持信息的连贯性、连续性。另一方面, 由于预测方程为微分方程, 从预测信息可以得到, 系统的长期发展趋势能够得到有效的反映, 但缺少波动性仍然是灰色预测方程的缺点。

表 2 预测结果

年份	产量/万吨	均值预测/万吨	相对误差/%
2001	45 263.7	44 975.5	-0.60
2002	45 705.8	45 680.2	0.01
2003	43 069.5	47 712.0	10.80
2004	46 946.9	48 844.7	4.00
2005	48 402.2	49 791.8	2.80

7 结束语

商空间理论讨论粒度信息的划分, 信息之间和信息粒度之间的拓扑关系, 从不同角度, 不同层次考察问题。这一思想是人类认识世界探求真理的必然路径。应用商空间理论的这一理念建立预测模型是必然的选择, 本文结合灰色预测理论的方法论建立商空间信息粒度的预测模型, 提出了在粒度集合上进行信息白化的概念, 其中半序结构的白化信息是建立预测方程的关键, 信息是逐渐增加的, 半序结构保证信息之间在拓扑关系上是相对紧密的。据此建立的方程在理论上是可靠的, 仿真实验也表明预测结果是可行的。

参考文献:

- [1] 张钹, 张铃. 问题求解的理论及应用[M]. 北京: 清华大学出版社, 1990.
- [2] 翁文波, 吕牛顿, 张清. 预测学[M]. 北京: 石油出版社, 1996.
- [3] 刘思峰, 党耀国, 方志国. 灰色系统理论及应用[M]. 北京: 科学出版社, 2004.
- [4] Yao Y Y, Zhong Ning. Granular computing using information table[M]// Data Mining, Rough Sets and Granular Computing Heidelberg: Physica Vedag, 2000: 102-124.
- [5] 张钹, 张铃. 商空间理论与粒度计算[J]. 计算机科学, 2003, 30(5): 1-3.
- [6] 莫旭, 扬风杰, 白洪涛, 等. 基于数据仓库的粮食预测系统[J]. 计算机工程, 2005, 131(17): 215-216.
- [7] 刘峰, 翟俊. 基于聚类分析和神经网络的时间序列预测方法[J]. 微电子学与计算机, 2006, 23(9): 85-87.