文章编号: 1001-0920(2004)08-0907-04

数据挖掘的系统云灰色预测方法研究

陈德军[1,2], 张玉民[2], 陈绵云[2]
(1. 武汉理工大学 信息工程学院, 湖北 武汉 430070; 2. 华中科技大学 控制科学与工程系, 湖北 武汉 430074)

摘 要: 剖析了系统云灰色预测模型的构造机理, 对其积分生成原理进行了论证, 并对其求解方法进行了深入研究, 提出了一种系统云灰色模型的解析预测公式. 结合数据库中"贫"信息和小样本序列数据的特点, 研究了用该模型进行数据挖掘的方法, 并用实例对解析预测公式和还原离散预测公式进行对比, 阐述了解析预测方法具有求解简单、结果详细、直观的优点.

关键词: 数据挖掘; 系统云; 灰色预测; 趋势关联
中图分类号: N941.5 **文献标识码**: A

On method of system cloud gray forecasting in data mining

CHEN De-jun, ZHANG Yu-min, CHEN Mian-yun
(1. College of Information Engineering, Wuhan University of Technology, Wuhan 430074, China; 2. Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China. Correspondent: CHEN De-jun, E-mail: mr-chendj@163.com)

Abstract: A constructing mechanism of the system cloud gray model is analyzed. The integral generating principle of the model is proved, and its solving method is researched deeply. A continuous forecasting expressions of system cloud gray model is proposed. According to the characteristics of "poor" information and small sample series in database, data mining method by system cloud gray model is investigated. The results are compared with continuous forecasting and discrete reductive forecasting by an example. The virtues of continuous forecasting expressions are clarified.

Key words: data mining; system cloud; gray forecasting; trend relation

1 引 言

在社会经济生态这类复杂系统中, 往往存在多种影响因素, 它们既相互独立, 又相互影响, 这种共同作用影响着系统的发展. 对此类系统的研究, 可采用灰色系统云模型[1](SCGM(1, h)). 系统云是指"贫"信息、多因子、不确定的错综复杂事物. SCGM(1, h)模型考虑了系统中各因素的耦合作用, 能合理反映系统的动态运动轨迹, 实现较为准确的预测. 研究其构造机理和简洁、直观的求解方法具有重要的理论和实用价值. 数据库中的时序数据非常普遍, 其时序数据的变化反映为一种趋势, 对时序数据进行数据挖掘已成为一类热点课题[2]. 而对于数据库中一类具有"贫"信息、小样本特点的数据序列, 可采用SCGM(1, h)模型实现预测. 可见, 对该模型及其在数据挖掘中的应用进行深入研究, 具有重要的理论和实际意义.

2 SCGM(1, h)模型的构造与求解

2.1 SCGM(1, h)模型的建模机理

本文仅考虑采用时序表征的系统行为. 记 $x_r^{(0)}$ 为参考时序, $x_c^{(0)}$ 为比较时序, 分别为

收稿日期: 2003-08-05; **修回日期**: 2003-10-15.
基金项目: 国家自然科学基金资助项目(79970025); 国防科研预研基金资助项目(00J15.3.3.JW0528).
作者简介: 陈德军(1964—), 男, 湖北天门人, 副教授, 博士, 从事灰色系统理论的研究; 陈绵云(1937—), 男, 湖北竹山人, 教授, 博士生导师, 从事灰色系统理论、决策支持系统等研究.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$x_r^{(0)} = \{x_r^{(0)}(1), x_r^{(0)}(2), \cdots, x_r^{(0)}(n)\},$$
$$x_c^{(0)} = \{x_c^{(0)}(1), x_c^{(0)}(2), \cdots, x_c^{(0)}(n)\}.$$

根据趋势关联函数[3]的定义, 可用趋势关联度

$$\Xi_{rc} = \frac{1}{n-1}\sum_{k=2}^{n}\xi_{rc}(k) \tag{1}$$

分析 $x_r^{(0)}$ 与 $x_c^{(0)}$ 的"相似性"和"接近性". 式中 $\xi_{rc}(k)$ 为趋势关联函数, 其定义见文献[3].

引理 1 令 $\{X_i^{(0)}(k) \mid k = 1, 2, \cdots, n;\ i = 1, 2, \cdots, h\}$ 是系统的观测时序, 相应有均值时序 $\{\overline{X}_i^{(0)}(k)\}$ 和均值累加生成时序 $\{\overline{X}_i^{(1)}(k)\}$, $k = 2, 3, \cdots, n$. 令 $s(f)$ 为已知的非齐次离散指数函数集, 设 $f_r \in s(f)$ 是定义在时域上的一个确定函数. 考虑一般广义能量系统的形式, 令 $f_r(k) = e^{A(k-1)}B - C$, 其中: $A \in R^{h\times h}, B \in R^{h\times 1}, C \in R^{h\times 1}$, $k = 1, 2, \cdots, n$. 若 $\{\overline{X}_i^{(1)}(k)\}$ 与非齐次离散指数函数 $f_r(k)$ 满意趋势关联, 则 SCGM $(1, h)$ 模型为

$$\dot{\hat{X}}^{(1)}(t) = \hat{A}\hat{X}^{(1)}(t) + \hat{U},\ t \geq 0, \tag{2}$$

其解为

$$\hat{X}^{(1)}(t) = e^{\hat{A}t}(\hat{X}^{(1)}(0) + \hat{A}^{-1}\hat{U}) - \hat{A}^{-1}\hat{U}, \tag{3}$$

离散形式为

$$\hat{X}^{(1)}(k) = e^{\hat{A}(k-1)}(\hat{X}^{(1)}(1) + \hat{A}^{-1}\hat{U}) - \hat{A}^{-1}\hat{U}. \tag{4}$$

式中

$$\overline{X}_i^{(0)}(k) = 0.5(X_i^{(0)}(k) + X_i^{(0)}(k-1)),$$
$$k = 2, 3, \cdots, n,\ i = 1, 2, \cdots, h;$$
$$\overline{X}_i^{(1)}(k) = \sum_{m=2}^{k}\overline{X}_i^{(0)}(m),$$
$$k = 2, 3, \cdots, n,\ i = 1, 2, \cdots, h;$$
$$\hat{X}^{(1)}(t) = [\hat{X}_1^{(1)}(t), \hat{X}_2^{(1)}(t), \cdots, \hat{X}_h^{(1)}(t)]^{\mathrm{T}},$$
$$\hat{X}^{(1)}(k) = [\hat{X}_1^{(1)}(k), \hat{X}_2^{(1)}(k), \cdots, \hat{X}_h^{(1)}(k)]^{\mathrm{T}}.$$

证明参见文献[1].

引理 1 根据趋势关联的原理, 提出了系统云模型, 揭示了系统的运动规律, 克服了用单独各种因素进行独立建模的缺点.

性质 1 模型(2)的累加生成过程是一种积分生成变换, 它有效地利用了已有灰色信息, 使系统动态模型更具合理性和可信性.

证明 由 Newton-Cotes 求积公式可知, 对于定积分 $I = \int f(x)\mathrm{d}x$, 可用被积函数 $f(x)$ 在 $[a, b]$ 上的一些点 $x_j\ (j = 1, 2, \cdots, n)$ 的值 $f(x_j) = f_j$ 的线性组合 $\sum_{j=0}^{n}A_j f_j$ 作为 I 的近似值, 即

$$\int f(x)\mathrm{d}x \approx (b-a)\sum C_j^{(n)} f(x_j) = \sum_{j=0}^{n}A_j f_j. \tag{5}$$

其中 $A_j = (b-a)C_j^{(n)}$ 称为 Cotes 系数, 它是不依赖于 $f(x)$ 的具体形式和积分区间 $[a, b]$ 的常数. 当 $n = 1$ 时, 可得梯形积分公式

$$\int f(x)\mathrm{d}x \approx \frac{(b-a)}{2}[f(a) + f(b)].$$

为了提高求积的精度, 可采用复化求积方法, 将 $[a, b]$ 等分成 n 个小区间, 在每个小区间上使用求积公式, 并将每个小区的计算结果累加起来, 得到 $[a, b]$ 上的积分近似值, 即

$$\int f(x)\mathrm{d}x = \sum_{j=0}^{n-1}\int_{j}^{j+1} f(x)\mathrm{d}x \approx \sum_{j=0}^{n-1}\frac{l}{2}[f(x_j) + f(x_{j+1})].$$

式中 $l = (b-a)/n$ 为每个小区间的长度.

在模型(2)中, 系统各因素的原始序列为

$$X_i^{(0)} = \{X_i^{(0)}(k) \mid X_i^{(0)}(k) \geq 0, k = 1, 2, \cdots, n\},$$

将原始序列的总区间 $[1, k]$ 划分为 $k-1$ 个小区间, 且在每个小区间上令

$$\overline{X}_i^{(0)}(k) = 0.5[X_i^{(0)}(k) + X_i^{(0)}(k-1)],$$

并取 $\overline{X}_i^{(1)}(k) = \sum_{m=2}^{k}\overline{X}_i^{(0)}(m)$, 由此得到的生成序列

$$\overline{X}_i^{(1)} = \{\overline{X}_i^{(1)}(k) \mid \overline{X}_i^{(1)}(k) \geq 0, k = 2, \cdots, n\}$$

即为积分生成序列, 记为

$$\overline{X}^{(1)}(k) = \mathrm{IGT}(\overline{X}^{(0)}(k)) = \sum_{m=2}^{k}\overline{X}^{(0)}(m).$$

2.2 预测模型求解

引理 2 SCGM $(1, h)$ 还原解(预测模型)的离散形式为

$$\hat{X}^{(0)}(k) = 2e^{\hat{A}(k-1)}(I + e^{-\hat{A}})^{-1}(I - e^{-\hat{A}})\hat{B}. \tag{6}$$

称此模型为离散预测模型, $\hat{B}$ 的表达式同模型(2).

证明参见文献[4].

引理 2 得到了 SCGM $(1, h)$ 预测模型的还原离散解, 为预测问题的求解提供了一种快速的求解形式, 但它不能提供预测结果的连续走势, 不能解决决策人员对预测趋势变化过程的直观需求.

定理 1 令 $\{X_i^{(0)}(k) \mid k = 1, 2, \cdots, n;\ i = 1, 2, \cdots, h\}$ 是系统的观测时序, 相应有均值时序 $\{\overline{X}_i^{(0)}(k)\}$ 和均值累加生成时序 $\{\overline{X}_i^{(1)}(k)\}$, $k = 2, 3, \cdots, n$. 若 $\{\overline{X}_i^{(0)}(k)\}$ 和非齐次离散指数函数 f_r 满意趋势关联, 则 SCGM $(1, h)$ 的预测模型为

$$\hat{X}^{(0)}(t) = e^{\hat{A}t}\hat{A}(\hat{X}^{(1)}(0) + \hat{A}^{-1}\hat{U}) = e^{\hat{A}t}\hat{A}\hat{B}, \tag{7}$$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

离散形式为

$$\hat{X}^{(0)}(k) = e^{\hat{A}(k-1)}\hat{A}(\hat{X}^{(1)}(0) + \hat{A}^{-1}\hat{U}) = e^{\hat{A}(k-1)}\hat{A}\hat{B}. \quad (8)$$

式中 $\hat{A}$, $\hat{B}$ 和 $\hat{U}$ 的求取同模型(2).

证明　在模型(2)中, 矢量 $\hat{X}^{(1)}(t) = [\hat{X}_1^{(1)}(t), \hat{X}_2^{(1)}(t), \cdots, \hat{X}_h^{(1)}(t)]^T$, 其中变量 $\hat{X}_i^{(1)}(t)$ 是对序列 $\overline{X}_i^{(1)} = \{\overline{X}_i^{(1)}(k) \mid \overline{X}_i^{(1)}(k) \geq 0, k = 2, \cdots, n\}$ 所代表曲线的描述. 由性质 1 可知, $\overline{X}_i^{(1)}$ 是 $\overline{X}_i^{(0)}$ 的积分生成, 若设描述原始序列的曲线为 $\hat{X}_i^{(0)}(t)$, 则有 $\hat{X}_i^{(0)}(t) = \hat{X}_i^{\circ(1)}(t)$; 若设 $\hat{X}^{(0)}(t) = [\hat{X}_1^{(0)}(t), \hat{X}_2^{(0)}(t), \cdots, \hat{X}_h^{(0)}(t)]^T$, 则由式(3)即可得式(7). □

通过式(7), 可得到一组连续的预测曲线, 它清楚地给出了预测问题的趋势, 并得出与引理 2 一致的结论.

推论 1　令 $\{X_i^{(0)}(k) \mid k = 1, 2, \cdots, n; i = 1, 2, \cdots, h\}$ 是系统的观测时序, 相应有均值时序 $\{\overline{X}_i^{(0)}(k)\}$ 和均值累加生成时序 $\{\overline{X}_i^{(1)}(k)\}$, $k = 2, 3, \cdots, n$. 若 $\{\overline{X}_i^{(0)}(k)\}$ 和离散齐次指数函数 $X(k) = e^{A(k-1)}\overline{B}$ 满意趋势关联, 则 SCGM(1, h) 的预测模型为式(7)和式(8).

证明　由定理 1 可知, 系统观测时序的预测序列为 $\hat{X}^{(0)}(k) = e^{\hat{A}(k-1)}\hat{A}\hat{B}$. 令指数函数 $X(k) = e^{A(k-1)}\overline{B} = e^{\hat{A}(k-1)}\hat{A}\hat{B}$, 即可得证. □

显然, 当关联因子数 $h = 1$ 时, 即可得 SCGM(1, h) 的预测模型.

3　SCGM(1, h)模型在数据挖掘中的应用

在数据库中, 存在一类"贫"信息的序列数据, 其主要特点如下[5]:

1) 样本数据量不大或有残缺(这主要由于许多现实条件、历史或环境的影响, 导致数据缺少且不精确), 数据输入失误(如空值数据);

2) 样本数据更新变换快(前期数据将对数据挖掘结果产生负面影响, 不应引入计算);

3) 整体数据规律相当复杂, 但在某一时间或空间的数据却有很强的规律性;

4) 数据概化以后出现的数据概念过于抽象, 描述信息过少.

具有以上特点的数据, 不具备大样本数据的特点, 但在这类数据的背后, 可能隐藏着某种规律, 如采用系统云灰色预测模型进行趋势预测, 可以得出有利于决策的结论.

3.1　用 SCGM(1, h)模型进行数据挖掘的方法

对于所要分析的数据, 往往来自于数据库和数据仓库, 为此必须先对要研究的数据序列进行查询, 然后使用模型求解. 具体方法如下:

1) 确定面向主题的数据集: 根据决策者的需求, 确定被预测系统包含的因子集 X, 进而进行数据清理、聚集和构造数据立方体.

2) 获得分析数据: 通过查询或概化, 得到所需因素集的序列数据 $X^{(0)} = \{x_i^{(0)}(k) \mid k = 1, 2, \cdots, n\}$, $i = 1, 2, \cdots, h$.

3) 根据定理 1, 计算累加序列的拟合序列 $\hat{X}^{(1)} = \{\hat{x}_i^{(1)}(k) \mid k = 1, 2, \cdots, n\}$, $i = 1, 2, \cdots, h$.

4) 求出序列趋势关联度, 确定预测模型的合理性: 以原始序列为参考序列, 以预测序列为比较序列, 按式(1)计算趋势关联度, 确定该模型是否可作为预测模型, 并获得预测结果.

3.2　实例分析

以我国国民经济增长速度预测为例, 按 3.1 节给出的步骤进行求解. 选定国民经济系统, 并确定 1994～1999 年国内生产总值($X_1^{(0)}$), 工业总产值($X_2^{(0)}$)和交通运输电信业总产值[6]($X_2^{(0)}$)为影响国家经济的代表因素(见表 1 中数据 $X_i^{(0)}(k)$), 对我国国民经济状况进行分析.

由文献[1]给出的参数计算公式, 可得式(7)的结构参数, 进而计算得到拟合序列 $\{\hat{X}_i^{(1)}(t)\}$.

$$\hat{A} = \begin{bmatrix} 12.399\,4 & -21.613\,5 & -58.363\,8 \\ 6.681\,6 & -11.386\,1 & -33.785\,5 \\ 0.265\,6 & -0.469\,1 & -1.143\,2 \end{bmatrix},$$

$$\hat{B} = \begin{bmatrix} 544\,557.8 \\ 232\,934.2 \\ 28\,593.6 \end{bmatrix}, \hat{U} = \begin{bmatrix} 44\,550.3 \\ 19\,290.8 \\ 2\,444.5 \end{bmatrix}.$$

由式(1), 计算得拟合序列与观测序列的积分生成序列 $\{\overline{X}_i^{(1)}(k)\}$ 的趋势关联度为 $\varepsilon_{11}^{(1)} = 0.993\,5$, $\varepsilon_{22}^{(1)} = 0.990\,3$, $\varepsilon_{33}^{(1)} = 0.990\,8$. 可见, 二者的趋势关联性很强, 可选用 SCGM(1, 3)模型来近似构造反映国民经济的系统模型.

由连续预测模型(7), 可得观测序列 $\{X_i^{(0)}(k) \mid k = 1, 2, \cdots, 6; i = 1, 2, 3\}$ 的预测曲线($\hat{X}_i^{(0)}(t)$ ($i = 1, 2, 3$). 图 1 中: $X_{ai}(k)$, $X_{bi}(k)$ 和 $X_{ci}(k)$ 分别表示原始值、还原解(预测模型)的离散公式(6)和连续模型(7)的预测值. 表 3 中: $\hat{X}_{bi}^{(0)}(k)$ 和 $\hat{X}_{ci}^{(0)}(k)$ 分别表示采用离散模型和连续模型所得的预测值. 由图 1 和表 3 知, 两种方法的预测效果基本一致, 但连续模型更为直观.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 1994～ 1999 年国内生产总值、工业生产总值、交通运输电信业总产值 (亿元)

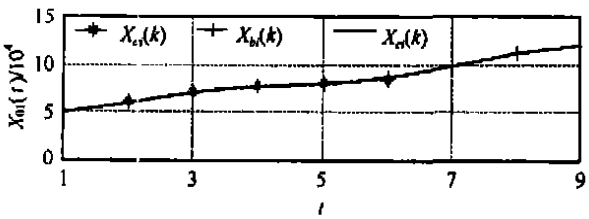
t	1994	1995	1996	1997	1998	1999
k	1	2	3	4	5	6
$X_1^{(0)}(k)$	46 759.4	584 78.1	67 884.6	74 462.6	78 345.2	81 910.9
$X_2^{(0)}(k)$	19 359.6	247 18.3	29 082.6	32 412.1	33 387.9	34 975.2
$X_3^{(0)}(k)$	2 685.9	3 054.7	3 494.0	3 797.2	4 121.3	4 459.5
$\overline{X}_1^{(1)}(k)$	46 759.4	99 378.15	162 559.5	233 733.1	310 137	390 265.05
$\overline{X}_2^{(1)}(k)$	19 359.6	41 398.55	68 299	99 046.35	131 946.35	166 127.9
$\overline{X}_3^{(1)}(k)$	2 685.9	5 556.2	8 830.55	12 476.15	16 435.4	20 725.8

表 2 积分累加序列的拟合值 (亿元)

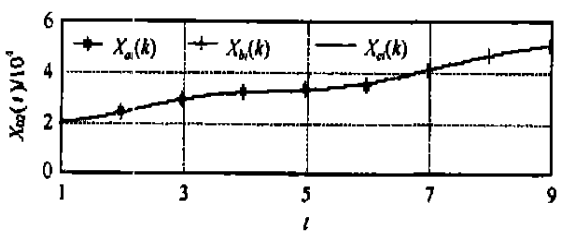
k	1	2	3	4	5	6
$\hat{X}_1^{(1)}(t)$	47 170.7	99 440.9	162 336.9	234 006.0	309 894.7	390 394.2
$\hat{X}_2^{(1)}(t)$	19 652.0	41 493.1	68 226.7	99 172.9	131 844.2	166 081.3
$\hat{X}_3^{(1)}(t)$	2 677.3	5 541.7	8 816.3	12 476.0	16 442.1	20 748.1

表 3 离散与连续模型的预测值 (亿元)

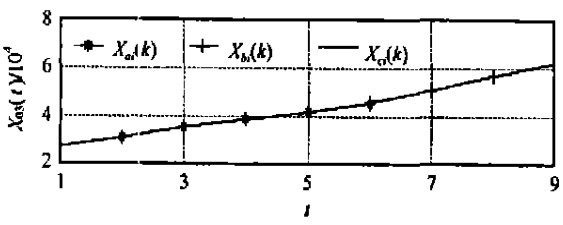
k	$\hat{X}_{b1}^{(0)}$	$\hat{X}_{b2}^{(0)}(k)$	$\hat{X}_{b3}^{(0)}(k)$	$\hat{X}_{c1}^{(0)}(k)$	$\hat{X}_{c2}^{(0)}(k)$	$\hat{X}_{c3}^{(0)}(k)$
7	95 992.6	40 841.2	5 049.3	95 991.3	40 820.1	5 047.2
8	108 140.9	46 57.1	5 617.7	108 453.0	46 525.4	5 621.9
9	117 933.5	50 583.0	6 165.8	117 953.4	50 623.2	6 163.1



(a) 国内生产总值



(b) 工业生产总值



(c) 交通运输电信生产总值

图 1 离散、连续预测求解模型的预测结果

参考文献(References):

[1] 陈绵云. 制定城市总体规划的灰色系统方法[J]. 华中理工大学学报, 1990, 18(3): 1-7.
(Chen M Y. Application of grey system method in overall urban planning[J]. *J of Huazhong University of Science and Technology*, 1990, 18(3): 1-7.)

[2] Han J, Camber M. 数据挖掘: 概念与技术[M]. 范明, 孟小峰译. 北京: 机械工业出版社, 2001.

[3] 陈绵云. 趋势关联度及其在灰色建模中的应用[J]. 华中理工大学学报, 1994, 22(8): 66-68.
(Chen M Y. Tendency correlation and its application in grey system modelling[J]. *J of Huazhong University of Science and Technology*, 1994, 22(8): 66-68.)

[4] 陈绵云, 尹平林, 熊和金. SCGM $(1,h)_c$ 残差修正模型及其在柳州市总体规划中的应用[J]. 华中理工大学学报, 1993, 21(3): 86-71.
(Chen M Y, Yin P L, Xiong H J. The SCGM $(1,h)_c$ forecasting model with residual error modification and its application in overall urban planning of Liuzhou City [J]. *J of Huazhong University of Science and Technology*, 1993, 21(3): 86-71.)

[5] Xiong Hejin, Chen Dejun, Chen Mianyun. Study on gray method of data mining [J]. *Advances in System Science and Applications*, 2003, 3(2): 24-29.

[6] 国家统计局. 中国统计年鉴 2000[M]. 北京: 中国统计出版社, 2000.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net