学术论文

基于灰色系统理论的时序数据挖掘技术

刘 斌[1,2]，刘思峰[1]，党耀国[1,2]
（1. 南京航空航天大学经济与管理学院，南京 210016；2. 河南农业大学，郑州 450002）

［摘要］ 阐述了嵌入知识的数据挖掘思想和数据挖掘技术现状，结合灰色系统理论首次提出了时序数据挖掘的灰色系统方法集（GDMS），并以灰色系统中的GM（1，1）模型为例，介绍了其具体算法。应用此算法对上海市2002～2005年的上网户数进行了预测。

［关键词］ 灰色系统；嵌入知识；GDMS；预测
［中图分类号］ O159；TP18 ［文献标识码］ A ［文章编号］ 1009－1742（2003）09－0032－04

1 前言

近年来，随着计算机技术和网络技术的迅猛发展，使得高效计算、大容量存储、联机分析处理及辅助决策成为可能。虽然人们探索了许多组织和应用数据的方法，但是面对日益膨胀的数据，人们往往处于一种尴尬的境地。由于缺乏有益的工具，被收集的数据已经远远超出了人类处理的能力，结果导致被收集的数据在大型数据库中成为“data tomb”，很少被访问，决策时常是在无法利用如此丰富的数据而不得不依靠决策制定者的直觉来做出。因此，利用新技术把数据转换成知识是大型数据信息系统面对的挑战。数据挖掘（data mining）这种致力于数据分析和理解、揭示数据内部隐藏知识的技术[1]，自然成为信息技术的一个新的热点。

数据挖掘工作是一个复杂而系统化的工程，它的发展是信息技术革命的必然趋势。在其发展过程中，产生了许多新概念和新技术，并且随着研究的深入，一些概念和技术趋于成熟，同时它的出现为许多技术和方法提供了丰富的土壤。灰色系统理论是近年来应用广泛的解决不确定问题的一门新兴横断学科，它主要通过对信息的生成、开发，提取有价值的信息，实现对系统运行行为的正确认识和有效控制。笔者从嵌入知识的数据挖掘思想出发，提出了时序数据挖掘的灰色系统方法集（GDMS），并以GM（1，1）模型为例，介绍了其具体算法。

2 嵌入知识的数据挖掘思想

数据挖掘出现在20世纪80年代后期，90年代有了突飞猛进地发展，至今已经是数据库研究、开发和应用最活跃的分支之一。简单地说，数据挖掘是从大量数据中提取或挖掘知识，是大型数据库中知识发现（KDD，knowledge discovery in database）的一个步骤。KDD主要利用某些特定的知识发现算法，在一定的运算效率的限制内，从数据库中发现有关的知识。它是一个多步骤的对大量数据进行分析的过程[1]，包括数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估和知识表示几个阶段。其中：数据清理实现清除不一致的数据；数据集成将多种数据源组合起来；数据选择阶段从数据库中检索与分析与任务相关的数据；数据变换将数据统一成适合挖掘的形式；数据

［收稿日期］ 2002－11－11；修回日期 2003－01－13
［基金项目］ 河南省自然科学基金资助项目（99401180）；南京航空航天大学特聘教授基金资助项目（1009－260812）
［作者简介］ 刘 斌（1971－），男，河南郑州市人，河南农业大学讲师，南京航空航天大学博士研究生

挖掘使用智能方法提取数据模式；在模式评估阶段，根据某种兴趣度度量，识别表示知识的真正有趣的模式；在知识表示阶段使用可视化和知识表示技术，向用户提供挖掘的知识。也就是说，首先从数据源中抽样和按照某种方式选出执行 KDD 过程的数据，通过预处理删除不合理数据，实现对数据进行合理的变换；然后通过数据挖掘，建立数学模型，进行解释或预测，形成 KDD 报告。

嵌入知识的数据挖掘技术是当今先进的建模思想和数据库中知识发现技术的体现[2]。传统的建模思想似乎更注重数据本身，如统计及假设检验方法等，更关心“采样”得到的数据中隐含的规律性。一般说来，关心数据本身建模并没有错，但是不考虑经验知识的数据建模无疑存在一个最大的缺陷，就是忽略了本该利用的信息，尤其是当人们有很多关于工艺过程经验时，这种做法损失更大。因此，把经验知识嵌入到数据的建模过程，是对于传统建模方法的有效补充和完善，图 1 所示的是嵌入知识的数据挖掘的建模过程。

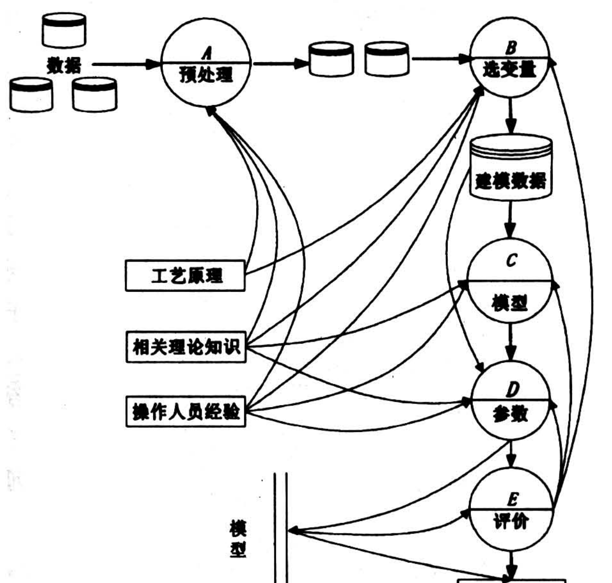

图 1　嵌入知识的数据挖掘建模过程

Fig.1　The data mining modeling process embedded with knowledge

目前，在数据挖掘中常用的技术有人工神经网络、决策树、遗传算法、近邻算法及规则推导等，应用这些技术可以完成对数据的特征化和区分、关联分析、分类和预测、聚类分析、孤立点分析、演变分析等挖掘功能[2]。鉴于数据、数据挖掘任务和数据挖掘方法的多样性，数据挖掘方法和技术的研究成为数据挖掘中最具挑战的课题，特别是对复杂数据类型的挖掘。仅仅依靠以往那种由简单汇总、按指定模式去分析的统计方法是无法完成数据分析的，研究和开发分析庞大数据资料的技术的任务日益紧迫，这就要求对数据挖掘技术的研究必须综合应用相关学科的知识和数据处理。基于这种思路，提出了基于灰色系统理论的时序数据挖掘技术。

3　基于灰色系统理论的时序数据挖掘技术

灰色系统理论的主要任务之一，就是根据社会、经济等系统的行为特征数据，寻求因素之间和因素本身的数学关系与变化规律。灰色系统利用对原始数据的整理来寻求其变化规律，这是一种就数据寻找数据的现实规律的途径。灰色系统理论认为，尽管客观系统表象复杂，数据离乱，但它总是有整体功能的，因此必然蕴涵某种内在规律。关键在于如何选择适当的方法去挖掘它和利用它，一切灰色序列都能通过某种生成，弱化其随机性，显示其规律性[3]。刘思峰教授提出的算子理论[4]成功地解决了系统数据预处理的难题，通过定性分析和定量分析相结合，设法排除系统行为数据所受到的冲击波干扰，实现正确反映系统的真实变化规律。

从嵌入知识的数据挖掘的观点看，灰色系统建模本身就是一种知识发现。鉴于经济现象的数据往往是时序数据，领会嵌入知识的数据挖掘思想，应用灰色系统的已有理论，可以提出基于灰色系统理论的时序数据挖掘方法集（GDMS），具体的主要方法如下：

灰色序列生成技术　通过对研究对象系统的分析，科学应用序列算子完成对数据的预处理；

灰色关联分析技术　依据所要分析序列的几何相似性，挖掘出它们之间的关联性；

灰色关联聚类分析技术　根据对数据序列的关联分析及临界值，划分对象子群；

灰色预测技术　通过对原始序列的预处理，挖掘系统潜在规律，利用灰色差分方程和灰色微分方程之间的互换，对离散的数据序列建立连续的动态微分方程，实现时间序列的预测。

笔者仅以灰色系统预测模型体系中的 GM(1，1) 为例，来描述具体的基于 GM(1，1) 的数

据挖掘技术。图2所示的就是该数据挖掘技术的建模过程。

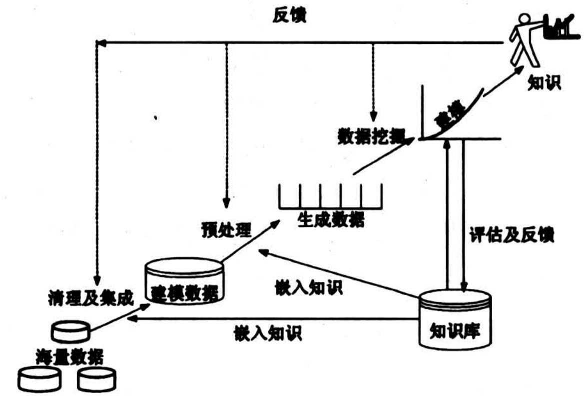

图2 基于GM（1，1）的数据挖掘建模过程

Fig. 2 The data mining modeling process based on GM (1, 1)

GDMS具体算法用伪代码描述如下：

Step 1 $p \leftarrow 0.98$ 设定所要建立模型的平均模拟精度 p（一般不低于98 %）。

Step 2 $X^{(0)} \leftarrow (x^{(0)}(1),\ x^{(0)}(2),\ \cdots,\ x^{(0)}(n))$ 输入原始数据。

Step 3 For $i \leftarrow 1$ to n,

$x^{(1)}(i) \leftarrow \sum_{k=1}^{i} x^{(0)}(k)$ 对原数据进行1-AGO[3,4]。

Next

Step 4 For $i \leftarrow 2$ to n,

$Z^{(1)}(i) = x^{(1)}(i) + x^{(1)}(i-1)$ 求出 $X^{(1)}$ 的均值生成。

Next

Step 5 确定 a，b，根据GM（1，1）的步骤进行建模。运用最小二乘法求出在 $Z^{(1)}$ 背景值下的灰色发展系数 $-a$ 和灰色作用量 b。（具体的GM（1，1）建模步骤可参考文献［3，4］）。

Step 6 For $k \leftarrow 2$ to n,

$\hat{x}^{(0)}(k) \leftarrow (x^{(0)}(1) - b/a)(e^{-a} - 1)e^{-a(k-2)}$，求出各数据的模拟值。

Next

Step 7 确定 ε 及 p'，求出平均相对模拟误差 ε 和模拟精度 p'。

Step 8 IF p', $\geqslant p$ Then

For $k \leftarrow n+1$ to $n+L$ 进行 L 步的预测。

$\hat{x}^{(0)}(k) \leftarrow (x^{(0)}(1) - b/a)(e^{-a} - 1)e^{-a(k-2)}$

Next

GOTO Step 9

Else

$X^{(0)} \leftarrow X^{(0)}D$，对 $X^{(0)}$ 施以某种缓冲算子，将对系统的定性分析融入到算子中。

GOTO Step 3

END IF

Step 9 OUTPUT a，b 及模拟值、模拟误差、平均相对误差及所需的预测值。

Step 10 END

灰色系统理论中包含了几种弱化和强化算子，以弱化或强化时序数据的增长趋势，用户可根据实际需要构建科学的算子，用于系统分析中的数据预处理。

4 实例

以上海市上网户数[5]为例来说明基于GM（1，1）的数据预测技术。

首先，基于预测的连贯性和相关性原则，以及灰色系统新息优先的建模思想，选取1996—2001年上网户数（单位为万户）作为原始数据：

$X = (0.33, 0.90, 10.24, 42.24, 88.24, 104.10)$。

可以发现，其增长势头很猛，每年均有近1～10倍的增长率，当然，在中国现有的经济情况下，如此高的增长率不可能一直保持。因此，用现有数据直接建模预测，预测结果是不能接受或得到认同的。经过认真分析，认识到增长速度高主要是由于基数低，基数低的原因则是Internet的刚刚兴起，人们接受新生事物需要一个过程。因此，要进行若干年后上网户数的预测，必须弱化其增长趋势，必须将对人们接受事物的过程这个现实考虑到序列中。为此利用弱化缓冲算子[4]。

设原始数据序列 $X = (x(1),\ x(2),\ \cdots,\ x(n))$：

令 $XD = (x(1)\mathrm{d},\ x(2)\mathrm{d},\ \ldots,\ x(n)\mathrm{d})$,

其中 $x(k)\mathrm{d} = \frac{1}{n-k+1}(x(k) + x(k+1) + \cdots. + x(n))$，$k = 1, 2, \cdots, n$，称 XD 为 X 的一阶缓冲算子；

令 $XD^2 = XDD = (x(1)\mathrm{d}^2,\ x(2)\mathrm{d}^2, \cdots,\ x(n)\mathrm{d}^2)$,

其中 $x(k)\mathrm{d}^2 = \frac{1}{n-k+1}(x(k)\mathrm{d} + x(k+1)\mathrm{d} + \cdots + x(n)\mathrm{d})$，$k = 1, 2, \cdots, n$，称 XD^2 为 X 的二阶缓

冲算子。

对原始序列采用一阶缓冲算子，得到 XD = (41.0085, 49.144, 61.205, 78.193, 96.17, 104.10)；对原始序列采用二阶缓冲算子，得到 XD^2 = (71.637, 77.762, 84.917, 92.821, 100.315, 104.10)。

根据上述的具体算法，先设定要求的模拟精度为 98 %；而后按 GM (1, 1) 的算法进行建模。经 3 次循环运算，可以得到满足精度要求的模型。具体的 3 次建模情况见表 1。

表 1　三种模型情况比较

Table 1　The comparing of three models

模型	预处理	ε / %	p' / %	p /%	建模要求
1	无	451.8	<0	98	不满足
2	XD	4.54	95.46	98	不满足
3	XD^2	1.31	98.69	98	满足

可见，采用合适预处理后的数据建模有利于取得较高的预测精度和模拟精度。运用 XD^2 进行建模其模拟精度满足要求。因此，对 XD^2 进行上网户数的预测，得到预测模型为：

$$\hat{x}(1996 + t) = 1\,040.930\,9e^{0.073135t} - 969.293\,9。$$

其平均模拟误差为：1.31 %，并可得 2002—2005 年的上网户数分别为：113.851，122.49，131.784，141.783。

4　结语

1）基于嵌入知识的数据挖掘思想，利用灰色系统理论的已有成果，首次提出了基于灰色系统理论的时序数据挖掘的技术集（GDMS），丰富了数据挖掘技术，特别是对时序数据的挖掘具有现实意义。当然，所有的数据挖掘技术，都要针对特定的对象，应用基于灰色系统的数据挖掘技术时，使用者应当注重系统信息和系统现象的分析，采用科学的和恰当的数据预处理。

2）应用基于灰色系统的数据挖掘技术时，如何用模型的形式来嵌入知识是一个具有复杂性的难点问题。钱学森教授等提出的“从定性到定量综合集成方法”[6, 7]可以对此问题提供有益的帮助和启示。

参考文献

[1]　Han Jiawei, Kamber M. 数据挖掘：概念与技术 [M]. 范　明等译. 北京：机械工业出版社，2001

[2]　王迎军. 供应链管理实用建模方法及数据挖掘 [M]. 北京：清华大学出版社，2001

[3]　Liu Sifeng, Lin Yi. An introduction to grey systems: foundations, methodologies and applications [M]. Slippery Rock, IIGSS Academic Publisher, 1988

[4]　刘思峰，郭天榜，党耀国，等. 灰色系统理论及其应用 [M]. 北京：科学出版社，1999

[5]　上海统计局. 上海统计年鉴 [M]. 北京：中国统计出版社，2002

[6]　于景元. 钱学森的现代科学技术体系与综合集成方法论 [J]. 中国工程科学，2001，(3)11：10～18

[7]　于景元，涂元季. 从定性到定量综合集成方法——案例研究 [J]. 系统工程理论与实践，2002，(5)：1～7，42

The Time Sequence Data Mining Techniques Based on Grey System Theory

Liu Bin[1,2], Liu Sifeng[1], Dang Yaoguo[1,2]

(1. *College of Economics and Management*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing* 210016, *China*; 2. *Henan Agricultural University*, *Zhengzhou* 450002, *China*)

[Abstract] This paper expatiates the thoughts of data mining with embedded knowledge and the techniques status quo of data mining. Based on the thoughts and the grey system (GS) theory, it proposes the GS-based data mining method set (GDMS) for time sequence first. Then this paper introduces the idiographic arithmetic with GM(1, 1) as an example. Last, it forecasts the total homes connecting with Internet in Shanghai in 2002—2005 by the arithmetic.

[Key words] grey system theories; embedded knowledge; time sequeke data mining; GDMS; forecast