

●丁 洁 (武汉大学信息管理学院 湖北 430072)

基于灰色灾变原理的互联网用户人数预测模型

摘要: 为了解决传统的 GM (1, 1) 模型在互联网发展预测中存在的历史数据跳变的问题, 本文根据灰色灾变原理, 基于互联网发展迅速、预测周期短等特点, 介绍了一种将 GM (1, 1) 模型与一元线性回归模型结合起来的方法。使用 1997 年 10 月至 2005 年 1 月的数据, 建立了我国互联网用户人数的预测模型, 并进行了检验, 证明该方法在实际应用中取得了很好的预测效果。

关键词: 因特网; 用户研究; 预测方法

Abstract: In the traditional GM (1, 1) forecasting model, there exists the problem of the aberrant points' presence of the history data. Based on the grey disaster theory and the characteristics of the Internet which is under rapid development and difficult to make a long term forecast, this paper presents a new method which combines the GM (1, 1) forecasting model with the subsection linear regression functions. A forecast model is established using the data of the Internet user growth of China from October, 1997 to January, 2005. The model proves to be very precise in practical application.

Keywords: Internet; user study; forecasting method

我国互联网用户人数、域名注册、上网计算机数等方面的统计信息对我国政府和企业掌握互联网的发展情况起着很重要的作用。而通过分析历史数据, 对未来进行准确预测, 并为政府和企业提供决策依据以制定出科学合理的互联网发展目标, 是一项十分有意义的工作。

在互联网发展预测中, 多采用数理统计回归分析的方法进行分析, 掌握的历史数据量愈大, 预测愈精确, 然而互联网在我国的发展不过数十年, 因此历史数据量较小。但是, 灰色预测 GM (1, 1) 模型因为要求样本数据量少且对短期预测精度高而使用广泛; 而线性回归分析模型主要适用于线性增长趋势的指标, 对序列数据存在的跳变问题无法预测。现将两种方法有机地结合起来, 即采用 GM (1, 1) 模型与线性回归组合的预测方法, 可以克服 GM (1, 1) 模型与一元线性回归方法各自的缺点。

本文根据新方法, 建立并检验了“我国互联网用户人数”预测模型。结果表明在实际应用中, 使用该方法建立的模型均比单独使用 GM (1, 1) 模型或线性回归模型的精度高。

1 GM (1, 1) 模型与一元线性回归组合模型

根据灰色灾变预测中灾变日期的预测原理, 将 GM (1, 1) 模型与一元线性回归方法结合起来, 在“跳变点”(跳变点的数据值相对于前连续的两个点来说显出明显的非线性特点, 即不在同一直线上) 采用跳变预测函数预测

其跳变值, 在非跳变点采用线性回归预测。

1) 按时间先后顺序列出样本序列:

$$x_{(0)} = (x_{(0)}(1), x_{(0)}(2), \dots, x_{(0)}(n)) \quad (1)$$

以及时间序列:

$$q_{(0)} = (q_{(0)}(1), q_{(0)}(2), \dots, q_{(0)}(n)) \quad (2)$$

2) 依原始序列画出折线图, 其中, $q_{(0)}(n)$ 为 X 轴, $x_{(0)}(n)$ 为 Y 轴。找出灾变点(跳变点)。根据灰色灾变原理, 将数据中跳变点的跳变值构成数据序列(跳变点是较前两个连续点, 显示出非线性特点, 明显不在同一直线上。一般来讲, 折线图的第一个点不看作跳变点)。

$$x_{A(0)} = (x_{A(0)}(1), x_{A(0)}(2), \dots, x_{A(0)}(n_A)) \subset x_{(0)} \quad (3)$$

其中: $x_{A(0)}$ 是跳变数据序列。

$$q_{A(0)} = (q_{A(0)}(1), q_{A(0)}(2), \dots, q_{A(0)}(n)) \subset q_{(0)} \quad (4)$$

其中: $q_{A(0)}$ 是对应的跳变时间的序列。

3) 用 GM (1, 1) 模型分别计算出跳变时间和相应值的函数并且预测出下一个或几个跳变时间和跳变值。

由 GM (1, 1) 模型计算 $q_{(0)}$ 和 $x_{A(0)}$ 的跳变函数 $\hat{q}_{A(0)}(n+1)$ 和 $\hat{x}_{A(0)}(n+1)$, 如下所示:

$$\begin{cases} \hat{q}_{(1)}(n+1) = \left[q_{(0)}(0) - \frac{\mu_q}{\alpha_q} \right] e^{-\alpha_q n} + \frac{\mu_q}{\alpha_q} \\ \hat{x}_{A(1)}(n+1) = \left[x_{(0)}(0) - \frac{\mu_x}{\alpha_x} \right] e^{-\alpha_x n} + \frac{\mu_x}{\alpha_x} \end{cases} \quad (5)$$

其中, α_p 为发展系数, μ_p 为灰作用量, 它们是微分方程的参数。分别根据 $x_{A(0)}$ 和 $q_{(0)}$, 使用最小二乘法估计得

到: $\hat{q}_{A(0)}(n+1)$ 和 $\hat{x}_{A(0)}(n+1)$ 的这两个参数值。

使用跳变函数预测跳变时间和跳变值, 分别为:

$$\begin{cases} \hat{q}_{A(0)}(n+1) = \hat{q}_{A(1)}(n+1) - \hat{q}_{A(1)}(n) \\ \hat{x}_{A(0)}(n+1) = \hat{x}_{A(1)}(n+1) - \hat{x}_{A(1)}(n) \end{cases} \quad (6)$$

4) 总结。

(1) 若预测日期是预测灾变点, 则用 GM(1, 1) 灾变函数进行预测:

$$\hat{x}_{A(0)}(n+1) = \hat{x}_{A(1)}(n+1) - \hat{x}_{A(1)}(n) \quad (7)$$

(2) 若预测日期不是预测灾变点, 则使用线性回归模型:

$$y = kx + b \quad (8)$$

2 我国上网用户人数预测模型的建立与分析

2.1 我国上网用户人数预测模型的建立

互联网的发展问题是近年来信息经济学中的热点问题。1997年11月, 中国互联网络信息中心(CNNIC)第一次发布《中国互联网络发展状况统计报告》受到了普遍关注, 于是CNNIC从1998年7月开始, 每隔半年就发布一次《中国互联网络发展状况统计报告》, 公布我国因特网上上网计算机数、用户人数、用户分布、信息流量分布、域名注册等方面情况的统计信息。本文选取其中的部分数据, 以我国上网用户人数为例(见表1), 用上述介绍的模型方法进行应用分析。

表1 1998年7月—2003年1月我国互联网用户人数*

(单位: 万)

序号	1	2	3	4	5
年/月	1998/7	1999/1	1999/7	2000/1	2000/7
数值	117.5	210	400	890	1690
序号	6	7	8	9	10
年/月	2001/1	2001/7	2002/1	2002/7	2003/1
数值	2250	2650	3370	4580	5910

* 数据来源:《中国互联网络发展状况统计报告》1998年7月—2005年1月。

注: 由于第1次统计与第2次统计间隔时间不足半年, 为了构成等差时间序列, 故以1998年7月数据为第1次。为了下文预测, 现只取10个样本点观察。

1) 我国互联网用户人数值序列 $q_{(0)}$ 和 $x_{(0)}$ 如下:

$$q_{(0)} = (1998/7, 1999/1, \dots, 2003/1) \quad (9)$$

$$x_{(0)} = (x_{(0)}(1998/7), x_{(0)}(1999/1), \dots, x_{(0)}(2003/1)) = (117.5, 210, \dots, 5910) \quad (10)$$

为方便计算起见, 以1998年7月为1, 转换序列:

$$q_{(0)} = (1, 2, \dots, 10) \quad (11)$$

2) 画出折线图(见图1), 找出跳变点。用Excel建立模型, 画出折线图。由图1可清晰地指出跳变点(画圈表示)。

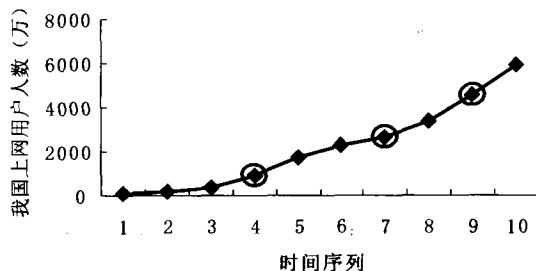


图1 我国1998年7月—2003年1月上网用户人数折线图

由图1观察出跳变点日期原始序列为:

$$\begin{aligned} q_{A(0)} &= (q_{A(0)}(1), q_{A(0)}(2), q_{A(0)}(3)) \\ &= (4, 7, 9) \end{aligned} \quad (12)$$

跳变原始序列为:

$$\begin{aligned} x_{A(0)} &= (x_{A(0)}(1), x_{A(0)}(2), x_{A(0)}(3)) \\ &= (890, 2650, 4580) \end{aligned} \quad (13)$$

3) 采用GM(1, 1)模型分别计算出跳变值和相应时间函数的参数并预测下一个跳变点。

由GM(1, 1)模型计算得:

$\hat{q}_{A(1)}(n+1)$ 的 α_p 为 -0.25 , μ_p 为 5.125 , $\hat{x}_{A(1)}(n+1)$ 的 α_p 为 -0.534 , μ_p 为 1467.411 , 从而得到预测方程如下:

$$\begin{cases} \hat{q}_{A(1)}(n+1) = 5.43e^{0.25n} \\ \hat{x}_{A(1)}(n+1) = 1505.21e^{0.534n} \end{cases} \quad (14)$$

根据方程组(14), 预测下面两个跳变点。

$n=3$ 时, 得:

$$\begin{cases} \hat{q}_{A(1)}(4) = 5.43e^{0.25 \times 3} \approx 12 \\ \hat{x}_{A(1)}(4) = 1505.21e^{0.534 \times 3} \approx 7470 \end{cases} \quad (15)$$

$n=4$ 时, 得:

$$\begin{cases} \hat{q}_{A(1)}(5) = 5.43e^{0.25 \times 4} = 14.76 \approx 15 \\ \hat{x}_{A(1)}(5) = 1505.21e^{0.534 \times 4} \approx 12740 \end{cases} \quad (16)$$

可知下两个跳变点分别是:

2004年1月($q_{A(1)}(4) \approx 12$), 2004年1月对应 $q_{(0)}$ 序列的12), 预测值为7470万人(如图2所示, 第12个点正好是跳变点, 预测值精度分析见后)。

2005年7月($q_{A(1)}(5) \approx 15$), 2005年7月对应 $q_{(0)}$ 序列的15), 预测值为12740万人(有待验证)。

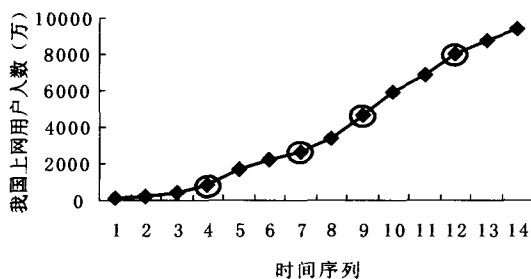


图2 我国1998年7月—2005年1月上网用户人数折线图

4) 计算跳变点之间的预测值。由于原始数据采用折线图反映历史数据的走势,根据分段函数的原理,对任意相邻两个跳变点来说,前一个跳变点的上一个点与后一个跳变点的上一个点之间可以认为是线性的关系。

由表1对应的序号和使用线性回归模型的判别方法可知2003年7月属于[2002年1月,2003年7月]区间(即时间序列的[8,11]),使用该区间数据建立线性回归模型:

由时序数据点(8,3370)和(10,5910),可得斜率 $k=1270$,则:

$$y = 1270x - 6790 \quad (17)$$

其中, x 为该年的时间序号。

2003年7月在 $q_{(0)}$ 中对应 $x=11$,代入方程得2003年7月,我国互联网用户数为7180万。

同理,可知2004年7月属于[2003年1月,2005年1月]区间(时间序列为[10,14]),使用该区间的数据建模,即由点(10,5910)和(12,7470)得:

$$y = 780x - 1890 \quad (18)$$

2004年7月,在 $q_{(0)}$ 中对应 $x=13$,代入方程得该年的互联网用户人数为8250万。

同理,预测2005年1月, $q_{(0)}$ 中对应的 $x=14$,代入方程(18),得9030万。

2.2 我国互联网用户人数预测模型的对比检验

下面就GM(1,1)模型与线性回归组合预测模型、线性回归模型以及GM(1,1)模型在该指标值的预测精度上进行比较。

本例线性回归模型函数:

$$y = 643.61x - 526.1 \quad (19)$$

表2中的预测精度显示,线性回归模型在短期的预测中是比较准确的,而随着预测周期的加大,预测精度会稍微下降,但总体来说还是能令人接受;而GM(1,1)模型,在短期预测中的精度就不太让人满意,而随着预测周期的加大,预测结果越来越不可信,这样的方法是不能用

表2 三种预测方法比较

(2003年7月—2005年1月我国互联网用户人数比较)

年份	实际值	组合预测模型		线性回归模型		GM(1,1)模型	
		预测值	精度(%)	预测值	精度(%)	预测值	精度(%)
2003/7	6800	7180	94.41	6553.61	96.38	7674.80	87.14
2004/1	7950	7470	93.96	7197.22	90.53	9943.14	74.93
2004/7	8700	8250	94.83	7840.83	90.12	12725.76	53.73
2005/1	9400	9030	96.06	8484.44	90.26	16267.20	26.95

作长期预测的;而本文所用的两种方法相结合的模型,无论在短期还是在长期的预测中,均令人满意,并且预测周期的加大对其精度也没有太大影响。试验证明,这种模型是实际应用中比较科学可靠的预测方法。

3 结论

本文根据灰色灾变原理,应用GM(1,1)模型和线性回归模型相结合的方法,对“我国互联网用户人数”建立预测模型。由折线图可知,该指标是一个呈指数型增长的曲线。从预测结果来看,通过与其他单一方法相对比,本文使用的综合方法比另外两种方法精度高,预测结果的精度基本不受预测周期长短影响。

然而在本方法的使用中,应注意以下几点:

1) 跳变点的选取至关重要。跳变点的定义本身是唯一的,但由于凭肉眼观察判断可能会造成误差,而导致跳变点选取错误。

2) 本方法对呈指数型增长以及呈线性增长趋势的指标,预测精度高;而对于其他曲线的指标,预测精度不够理想。□

参考文献

- 1 许秀莉,罗键.一种新的组合灰色神经网络预测模型.厦门大学学报,2002(3):164~167
- 2 鲍一丹等.基于GM(1,1)模型和线性回归的组合预测新方法.系统工程理论与实践,2004(3):95~98
- 3 邓聚龙.灰预测与灰决策.武汉:华中理工大学出版社,2002
- 4 傅立.灰色系统理论及其应用.北京:科学技术文献出版社,1992
- 5 布洛克威尔等著.时间序列的理论与方法(第二版).田铮译.北京:高等教育出版社,2001
- 6 陈业华.灰色灾变预测模型及其应用.北京航空航天大学学报,1998(1):79~82

作者简介:丁洁,女,1983年生。

收稿日期:2005-04-29