The Verhulst model with remedy and its application in forecasting quantity of student taking entrance examination to college*

LIU Bin[1], BI Qing-sheng[2]

(1. Department of Management Science, Henan Agricultural University, Zhengzhou 450002, China;

2. Ministry of Informationization Management, Henan Agricultural University; Zhengzhou 450002, China)

Abstract: The Verhulst model can be used to forecast the sequence, which is characterized as non-monotone and fluctuant sequence or saturated S-form sequence. According to the situation of national enrollment scale of college, this paper forecasts the quantity of students taking entrance examination to college with a Verhulst model with remedy based on data mining theories, and by the above model, some countermeasures are developed to the higher education of Henan province in China.

Key words: forecast; grey system; Verhulst model with remedy

1. Introduction

Grey system theory, which is introduced by professor DENG Ju-long in 1982, has been applied to system analysis and system modeling in some related fields, such as economy, management and education. Grey incidence analysis, grey cluster analysis and grey forecasting are the main parts in grey system and they can be used to analyze the system critical factors, class the research objects, and analyze the future, respectively. Generally, grey forecasting can find out and hold the development rules of object system with the processing the original data and the modeling grey models, and forecast the future of system with a scientific quantitative method. At present, GM(1,1), grey Markov model and Verhulst model are applied widely in some fields.

The Verhulst model can be used to forecast the original sequences with non-monotone wave-type characteristics, and can depict some process in saturation situation, such as S-type process, which is often applied to forecast the population, the growth of crops and the life-cycle of products. Undoubtedly, the quantity of students taking entrance examination to colleges belongs to this S-type process. In Henan province, as the biggest region of China in population, more and more people look forward to ensuring their ideals, and more and more students take part in the entrance examination to colleges. Obviously, the quantity of taking entrance examination to colleges increases and takes on a saturated situation. For knowing the quantity well and making countermeasures, the authors forecast the quantity with a Verhulst model with remedy, which will integrate the Verhulst model and the error remedy model well, and its modeling thought is shown as following Figure 1.

* This paper partly supported by Financial Program for Core Teacher of Henan's University.

LIU Bin, Ph.D., associate professor, head of Department of Management Science, Henan Agricultural University; research fields: include supply chain management, grey system theories and application.

BI Qing-sheng, associate professor, head of Ministry of Informationization Management, Henan Agricultural University; research fields: grey system theories, education management.

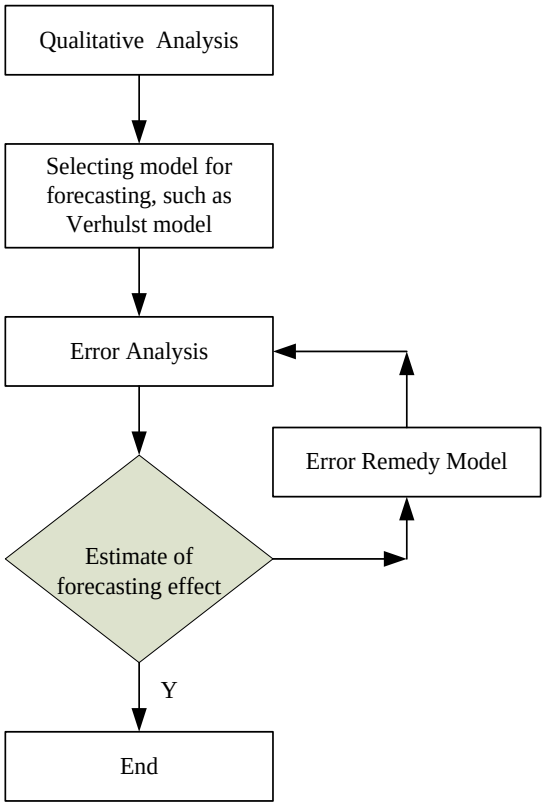

Figure 1 The modeling thoughts of the optimal Verhulst model

The rest of this paper is organized as follows: Section 2 is the modeling thought with grey system theories, and section 3 is the Verhulst model. Then, section 4 will introduce the Verhulst model with remedy, and section 5 is a case of the Verhulst model with remedy. Last, conclusions and countermeasures are developed.

2. Data mining with grey system theories

Data mining (DM) came forth in the late of 1980s and rapid developed in 1990s. Now it has already become one of most active sub-branches in studying, developing and applying of database. In short, data mining is defined to picking up or discovering knowledge from a great deal of data. It is a step of KDD (knowledge discovery in database). KDD is defined as utilize some specifically knowledge to discover arithmetic, and dig out involved knowledge in database with definite operation efficiency. KDD is a multi-step process of analyzing a great deal of data and it consists of data cleaning, data integration, data selection, data transformation, data mining, mode evaluating and knowledge expressing. Concretely, data cleaning can eliminate conflicting data; data integration will combine kinds of data source; data selection searches and analyses data related with tasks from database; data transformation unites data into a suitable form to mine; data mining picks up data pattern with intellectualized means; based on certain interesting degree, data evaluation recognizes really interesting pattern denoted knowledge; knowledge expressing, with visible techniques and knowledge expressing techniques, provides knowledge mined from database or data resource to user. In the other words, first we should sample from data source and select data in the light of certain data mode to carry out KDD. Then we can realize rational data

transforming with pretreatment to eliminate illogical or disorder data. Last, after establishing mathematic models to explain or predict via data mining, we can get the report of KDD.

The data mining techniques embedded knowledge embodies the present advanced thoughts of modeling and the KDD techniques of database (WANG, 2001). Traditional thoughts of modeling is seemed to pay more attention to data itself, for example, statistical method, hypothesis verifying method, etc., namely, they pay more attention to the rule characteristics hided in sampled data. Generally, it is not an error paying more attention to these data, but it is undoubted that there is being the biggest bug in establishment mathematic models without considering experience. Namely, we neglect the information that ought to be utilized. Especially if we have more experience and knowledge about the arts and crafts processing we will get more loss. Thus, it is effective makeup and improvement for traditional establishment method to embed experience knowledge to establishment process of data. Figure 2 will show the modeling process of data mining embedded knowledge. Present there are some kinds of techniques applied in data mining, for example, artificial neural network, decision tree, genetic algorithm and rule inferring, etc. We can apply these techniques to realize some data mining functions including data characterization and distinguishing, association analysis, classification and prediction, cluster analysis, outlier analysis, evolution analysis. Whereas the multiform data, data mining tasks and models, the study on data mining methods and techniques becomes the most challenging problem in data mining field, especially in complex data patterns. Only depending on hackneyed statistical methods, for example, simple gathering and analysis with appointed mode, we cannot complete those tasks of data mining. So it is urgent to study and develop analysis techniques applied in huge data information. Certainly, this task requests us to synthetically apply relative knowledge of different disciplines. Based on these thoughts, we provide the time sequence data mining techniques based on grey system theory (LIU, ZHANG & LIU, 2008).

One of the main tasks facing the theory of grey system is to seek the mathematic relations and movement rule among factors themselves and between factors, based on behavioral data of social, economic, etc. (DENG, 1982). In the GST, it is through the organization of raw data to sort out development laws. This is a path of finding out realistic governing laws from the available data. It is believed in the GST that even though objective systems phenomena can be complicated and related data chaotic, they always represent a whole, hence, implicitly contain some governing laws. The key for us to uncover and to make use of all these laws is how to choose appropriate methods (YU, 2001; YU & TU, 2002). The randomness of all grey sequence can be weakened to show its regularities through some generations. The operator theory provided by professor LIU Si-feng, is succeed to solve the difficult problem of data pretreatment. The purpose of introducing sequence operators is to eliminate the shock waves that system behavioral data was interfered in order to show the true face of the data collected, based on conclusions of qualitative analysis. Thus, in the view of DM techniques embedded knowledge, the modeling of grey system itself is a kind of KDD and the data of economic phenomena are often regard as time sequence data (WANG, 2002).

This paper only gives a forecasting demonstration with data mining modeling, which is depicted by the following Figure 2. Further, based on the problem focused by this paper, the Verhulst forecasting model will be

introduced in section 3.

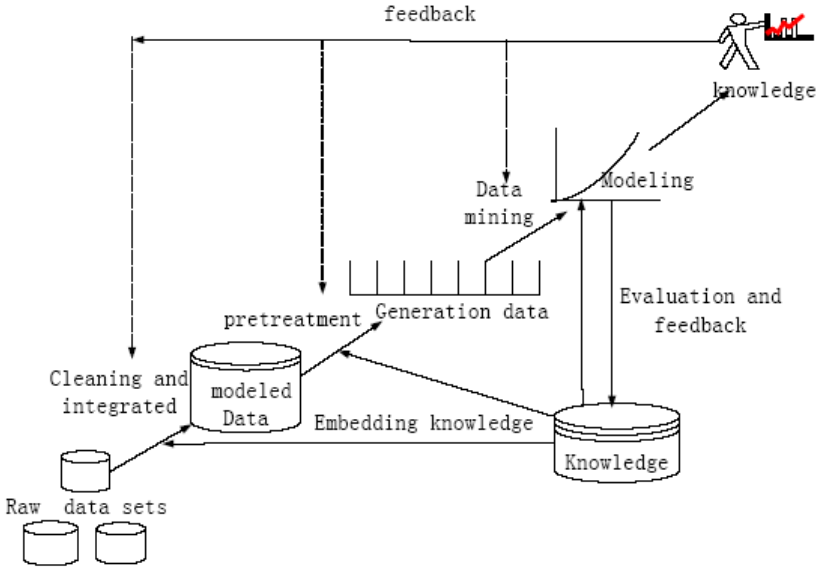

Figure 2 The modeling process for forecasting with data mining

3. Verhulst model

The basic principle and computing method are listed as following (LIU & LIN, 1998).

Definition 1: For the original data sequence $X^{(0)} = \{X^{(0)}(t) \mid t = 1,2,...,n\}$ and $X^{(1)}(k) = \sum_{t=1}^{k} X^{(0)}(t)$ is the 1-AGO sequence of $X^{(0)}$ and $Z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(1)}(k-1)}{2}$ where $k = 2,3,...,n.$ is a sequence mean generated of consecutive neighbor of $X^{(1)}$, then $X^{(0)} + a \cdot Z^{(1)} = b \cdot [Z^{(1)}]^r$ is called the GM(1,1) power model.

Definition 2: The equation $\frac{dx^{(1)}}{dt} + a \cdot x^{(1)} = b[x^{(1)}]^r$ is called the whitenization equation of the GM(1,1) power model.

Theorem 1: The solution of the whitenization equation of the GM(1,1) power model is given by

$$x^{(1)}(t) = \left\{e^{-(1-r)a \cdot t}\left[(1-r)\int b \cdot e^{-(1-r)a \cdot t}\,dt + C\right]\right\}^{\frac{1}{1-r}}$$

Theorem 2: Assume that $X^{(0)}$, $X^{(1)}$ and $Z^{(1)}$ are defined the same as in definition 1 and 2, and

$$B = \begin{bmatrix} -z^{(1)}(2) & \left[z^{(1)}(2)\right]^r \\ -z^{(1)}(3) & \left[z^{(1)}(3)\right]^r \\ ..... & ..... \\ -z^{(1)}(n) & \left[z^{(1)}(n)\right]^r \end{bmatrix}, \quad Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ ..... \\ x^{(0)}(n) \end{bmatrix}$$

Then the least square estimate of the parameter sequence $\hat{a} = [a\ b]^T$ of the GM(1,1) power model is given by

$$\hat{a} = \left[B^T B\right]^{-1} B^T Y.$$

Definition 3: When $r = 2$, $X^{(0)} + a \cdot Z^{(1)} = b \cdot [Z^{(1)}]^2$ is called the grey Verhulst model.

Definition 4: The equation $\frac{dx^{(1)}}{dt} + a \cdot x^{(1)} = b \cdot \left[x^{(1)}\right]^2$ is called the whitenization equation of the grey Verhulst model.

Theorem 3: The solution of Verhulst whitenization equation is given by

$$x^{(1)}(t)=\frac{a\cdot x^{(1)}(0)}{b\cdot x^{(1)}(0)+[a-b\cdot x^{(1)}(0)]e^{at}}$$

and the time response sequence of the grey Verhulst model is given by

$$\hat{x}^{(1)}(k+1)=\frac{a\cdot x^{(1)}(0)}{b\cdot x^{(1)}(0)+[a-b\cdot x^{(1)}(0)]e^{ak}}$$

From the solution of the Verhulst equation, it can be seen that when $t\to\infty$, if $a>0$, then $x^{(1)}(t)\to 0$; if $a<0$, then $x^{(1)}(t)\to\frac{a}{b}$. That is, when t is sufficient large, for any $k>t$, $x^{(1)}(k+1)$ and $x^{(1)}(k)$ will be sufficiently close. At this time, $x^{(0)}(k)=x^{(1)}(k)-x^{(1)}(k-1)>0$. So, the system approaches extinction.

When revolving practical problems, we often face with processes with the sigmoid sequence of raw data. In this case, we can take the sequences of the original data as $X^{(1)}$ and the 1-AGO sequence as $X^{(0)}$ to establish a Verhulst model to simulate $X^{(1)}$ directly. However, in practical management, when the accuracy of a Verhulst model is not meeting the requirements one can establish a GM(1,1) model or a linear model using the error sequence to remedy the original model in order to improve the accuracy. Here we will introduce a method to remedy the original model with a GM(1,1) model as an example.

4. The optimal Verhulst model with an error remedy model

Due to the fact that the restored values through derivatives and through inverse accumulating are not the same, in order to reduce possible errors caused in reciprocating operations, we often use the errors of $X^{(1)}$ to improve the simulation values of $X^{(1)}$

$$\hat{x}^{(1)}(k+1)=\left[x^{(0)}(1)-\frac{b}{a}\right]\cdot e^{-a\cdot k}+\frac{b}{a}$$

Definition 4: Assume that $\varepsilon^{(0)}=\left(\varepsilon^{(0)}(1),\varepsilon^{(0)}(2),...,\varepsilon^{(0)}(n)\right)$, where $\varepsilon^{(0)}(k)=x^{(1)}(k)-\hat{x}^{(1)}(k)$ is the error sequence of $X^{(1)}$. If there exists k_0 satisfying:

For any $k\ge k_0$, $\varepsilon^{(0)}(k)$ has the same sign;

$$n-k_0\ge 4,\ \left(\left|\varepsilon^{(0)}(k_0)\right|,\left|\varepsilon^{(0)}(k_0+1)\right|,...,\left|\varepsilon^{(0)}(n)\right|\right)$$

is called the error sequence of modellability, which is still denoted as

$$\varepsilon^{(0)}=\left(\varepsilon^{(0)}(k_0),\varepsilon^{(0)}(k_0+1),...,\varepsilon^{(0)}(n)\right)$$

Theorem 4: Assume that $\varepsilon^{(0)}=\left(\varepsilon^{(0)}(k_0),\varepsilon^{(0)}(k_0+1),...,\varepsilon^{(0)}(n)\right)$ is an error sequence of modellability with $\varepsilon^{(1)}=\left(\varepsilon^{(1)}(k_0),\varepsilon^{(1)}(k_0+1),...,...\varepsilon^{(1)}(n)\right)$, beginning its 1-AGO sequence, whose GM(1,1) time response sequence is given as $\hat{\varepsilon}^{(1)}(k+1)=\left[\varepsilon^{(0)}(k_0)-\frac{b_\varepsilon}{a_\varepsilon}\right]\cdot e^{-a_\varepsilon(k-k_0)}$, $k\ge k_0$, then the simulation sequence of the error sequence $\varepsilon^{(0)}$ is given by $\hat{\varepsilon}^{(0)}=\left(\hat{\varepsilon}^{(0)}(k_0),\hat{\varepsilon}^{(0)}(k_0+1),...,\hat{\varepsilon}^{(0)}(n)\right)$, where $\hat{\varepsilon}^{(0)}(k+1)=(-a_\varepsilon)\cdot\left[\varepsilon^{(0)}(k_0)-\frac{b_\varepsilon}{a_\varepsilon}\right]\cdot e^{-a_\varepsilon\cdot(k-k_0)}$, $k\ge k_0$.

Definition 5: if $\varepsilon^{(0)}$ is used to modify $\hat{X}^{(1)}$, the time response sequence after modification

$$\hat{x}^{(1)}(k+1)=\begin{cases}\left[x^{(0)}-\dfrac{b}{a}\right]\cdot e^{-a\cdot k}+\dfrac{b}{a} & k<k_0\\ \left[x^{(0)}-\dfrac{b}{a}\right]\cdot e^{-a\cdot k}+\dfrac{b}{a}\pm a_\varepsilon\cdot\left[\varepsilon^{(0)}(k_0)-\dfrac{b_\varepsilon}{a_\varepsilon}\right]\cdot e^{-a_\varepsilon(k-k_0)} & k\geq k_0\end{cases}$$

is called the GM(1,1) model with error modification, or remnant GM(1,1) for short.

Here, the sign of

$$\hat{\varepsilon}^{(0)}(k+1)=a_\varepsilon\cdot\left[\varepsilon^{(0)}(k_0)-\frac{b_\varepsilon}{a_\varepsilon}\right]\cdot e^{-a_\varepsilon(k-k_0)}$$

the error modification value, needs to be the same as the error $\varepsilon^{(0)}$.

It should be remarked that it is not necessary to adopt the GM(1,1) model as a remedy model, and the decision-maker should select the optimal method according to the characteristics of original sequences.

5. The application case

According to the statistics almanac of Henan province, we can get the quantity of students taking entrance examination to college from 2000 to 2008 (see Table 1). By the qualitative analysis, we adopt the above Verhulst model to depict this S-type original data, and Table 2 shows the related precision indices.

Table 1 The quantity of students taking entrance examination to college (Unit: thousand)

2000	2001	2002	2003	2004	2005	2006	2007	2008
269	291	355	498	596	722	784	791	905

Table 2 The simulation values, errors and relative errors

Year	Original value	Simulation value	Errors	Relative errors
2000	269	269	0	0
2001	291	331	40	12.08%
2002	355	401	46	11.47%
2003	498	476	-22	4.62%
2004	596	554	-42	7.58%
2005	722	632	-90	14.24%
2006	784	706	-78	11.04%
2007	791	774	-17	2.20%
2008	905	834	-71	8.51%

We can get the Verhulst model with the original sequence data,

$$\hat{x}^{(1)}(k+1)=\frac{77.337231}{0.071016+0.216483\cdot e^{-0.287499\cdot k}}$$

and the corresponding estimate indices are listed in Table 2.

We can get the error sequence $(0, 40, 46, -22, -42, -90, -78, -17, -71)$, and find that when $k\geq 4$, all data are negative. By analysis and simulation, we find that GM(1,1) can not satisfy the forecasting precision, and the six-order regression model can depict this sequence well. So, the remedy model is

$$R(k) = 0.1218k^6 - 3.2435k^5 + 32.551k^4 - 152.86k^3 + 337.44x^2 - 295.44x + 81.11,$$

$$R^2 = 0.9852$$

Then, the estimate of optimal Verhulst model with a six-order regression model is listed as in Table 3. And we can find the forecasting precision satisfies the requirement well.

Table 3 The simulation values, errors and relative errors

Year	Original value	Simulation value	Errors	Relative Errors
2000	269	269	0	0
2001	291	289	-2	0.67%
2002	355	359	-4	1.13%
2003	498	502	4	0.80%
2004	596	598	2	0.34%
2005	722	718	-4	0.55%
2006	784	794	10	1.28%
2007	791	799	8	1.01%
2008	905	925	20	2.21%

The optimal Verhulst model with remedy is

$$\hat{x}^{(1)}(k) = \begin{cases} \dfrac{77.337231}{0.071016 + 0.216483e^{-0.28749912\cdot(k-1)}} - R(k) & k < 4 \\ \dfrac{77.337231}{0.071016 + 0.216483e^{-0.28749912\cdot(k-1)}} + R(k) & k \geq 4 \end{cases}$$

where $R(k) = 0.1218k^6 - 3.2435k^5 + 32.551k^4 - 152.86k^3 + 337.44x^2 - 295.44x + 81.11$, and by this model we can get the quantity of students taking entrance examination to college from 2009 to 2010 are, respectively, 971 and 929.

6. Conclusions and countermeasures

With a Verhulst model with remedy, the authors forecast the future quantity of students taking entrance examination to college. From the simulation results, it can be found that its precision is very high, and the forecasting method integrates the qualitative analysis and quantitative modeling. It is interesting that the quantity of students taking entrance examination to college will reach into one million, and the authors think it will bring out a series of social problems if none improved policies are adopted by Chinese government. Thus, some countermeasures are developed as follows:

(1) The ministry of education should design a reasonable mechanism to allocate the enrollment quantity according to the quantity of students taking entrance examination to college;

(2) The examination method will be improved, and an united examination for whole nation should be adopted, which can depict the justice well;

(3) The ministry should encourage the local governments set up more colleges or universities in the middle and western regions.

Certainly, the education reforms will touch the behalf of some people in some cities, such as Beijing and Shanghai, however, it will benefit most of provinces in China. We think it is one of most important parts in our harmonious nation.

References:

DENG, J. L. (1982). *Grey systems: Society and economics*. Beijing: Press of National Defense Industry.

FAN, M. (2001). *Data mining: Conception and techniques*. Beijing: Mechanic Press.

LIU, S. F. & LIN, Y. (1998). *An introduction to grey systems theory*. IIGSS Academic Publisher.

LIU, B., ZHANG, R. & LIU, S. F. (2008). RFID system and its perspective analysis with KERGM(1,1) model. *Journal of Computers, 3*(7), 9-15.

WANG, Y. J. (2002). *Practical modeling methods for SCM and data mining.* Beijing: Tinghua University Press.

YU, J. Y. (2001). QIAN Xue-sen's contemporary system of science and technology and meta-synthesis. *Chinese Engineering Science, 3*(11), 10-18.

YU, J. Y. & TU, Y. J. (2002). Meta-synthesis-study of case. *System Engineering Theory and Practice, 22*(5), 1-7, 42.

(Edited by Nicole and Lily)