UDC 004.832.2

Data Mining Techniques Based on Grey System Theories for Time Sequence Data

Liu Bin[1,2], Zhang Hui[1],Liu Sifeng[2] and Dang Yaoguo[2]

[1]College of Information and Management Science
Henan Agricultural University, Zhengzhou, 450002 China
[2]College of Economics and Management
Nanjing University of Aeronautics and Astronautics, Nanjing, 210016 China
Liubhnau@163.com, huizi@henau.edu.cn, Sfliu@nuaa.edu.cn,
Iamdangyg@163.com

Abstract. Data mining is an interesting focus in computer science field now. This paper deals with data mining techniques based on Grey system theories for time sequence data. Firstly, thoughts of data mining with embedded knowledge are expatiated, and the status quo of Data mining techniques is presented briefly. Then, based on the above thoughts and the Grey system theories, data mining techniques based on Grey system theories for time sequence data are proposed for the first time, and the idiographic arithmetic with GM(1,1) as an example is introduced in this paper. Last, it forecasts the total homes in 2002~2005 connecting with Internet in ShangHai City by the arithmetic.

1. Introduction

With the rapid development of computer technologies and network technologies recent years, it has been likely to be true in computing with high effect, storing with large capability, OLAP (On-Line Analytical Processing) and the assistant decision-making. Though a great deal of methods have been researched in organizing and applying data, affronting the data expanding with the each day, we often be in the embarrassed situation. On the one hand, we can not deal with collected data because of the lack in appropriate tools, and on the other hand, these great deal of data in large scale database have become “Data Tomb”, which little are visited or utilized. Hence, decision-making often had to be made by the gnosis of decision-maker who cannot be able to utilize this abundant and ample data effectively. So, it is a challenge for big database to utilize new techniques to realize the transition from data to knowledge. Data mining, as the technology that takes up with data analyzing and comprehending and discovering the knowledge hided inside data [J.W.Han,2001], naturally becomes a new focus in information science fields.

Data mining is very complex and its development is the necessary trend. During its development some new concepts and new techniques have come into being, and with studying in depth some concepts and techniques will go to mature. At the same time, its emergence will become the foundation of some techniques and methods. Grey system theories is a new crossing discipline widely applied to solve the uncertainty questions recent years, and it is mostly through the generation and development of information and picks up the valuable information. Furthermore, it realizes to understand rightly and control effectively the action of system. This paper provides data mining methods for time sequence based on Grey system first.

The rest of this paper is organized as follows. In section 2, thoughts of data mining with embedded knowledge is presented. Then, data mining techniques based on Grey system theories for time sequence data is developed first and we design the idiographic arithmetic with GM (1,1) as an example in section 3. In section 4, we illustrate the above arithmetic. Last is the conclusion.

2. Thoughts of Data Mining with Embedded Knowledge

Data mining came forth in the late of 1980s and rapidly developed in 1990s, and 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 pick up or discover knowledge from a great deal of data. In short, it is a step of KDD (Knowledge Discovery in Database), which is defined to utilize some specifically knowledge to discover arithmetic, and digs out the involved knowledge in database with definite operation efficiency. In other words, KDD is a multi-step process of analyzing a great deal of data and consists of data cleaning, data integration, data selection, data transformation, data mining, mode evaluating and knowledge expressing [J.W.Han,2001]. Concretely, data cleaning can eliminate conflicting data; data integration will combine kinds of data source; data selection can search and analyze the data related with tasks from database; data transformation will unite data into a suitable form to mine; data mining will pick up data patterns with intellectualized means; based on certain interesting degree, data evaluation can recognize really the interesting pattern to denote knowledge; knowledge expressing, with visible techniques and knowledge expressing techniques, will provide the knowledge mined from database or data resource to user [J.W.Han,2001] [Y.J.Wang,2001]. Summarily, first we sample from data source and select data in the light of certain data mode to carry out KDD, and then realize the rational data transforming with pretreatment to eliminate the illogical or disorder data. Last, after establishing mathematic models to explain or predict via data mining, we can get the report of KDD.

Data mining techniques with embedded knowledge embodies the present advanced thoughts of modeling and the KDD techniques of database. Traditional thoughts of modeling is seemed to only pay more attention to data itself, for example, statistical method, hypothesis verifying method et al., namely, they only pay more attention to the rule characteristics hided in sampled data. Generally, it is 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 the experience knowledge to the establishment process of data. Fig. one will show modeling processes of data mining with embedded knowledge.

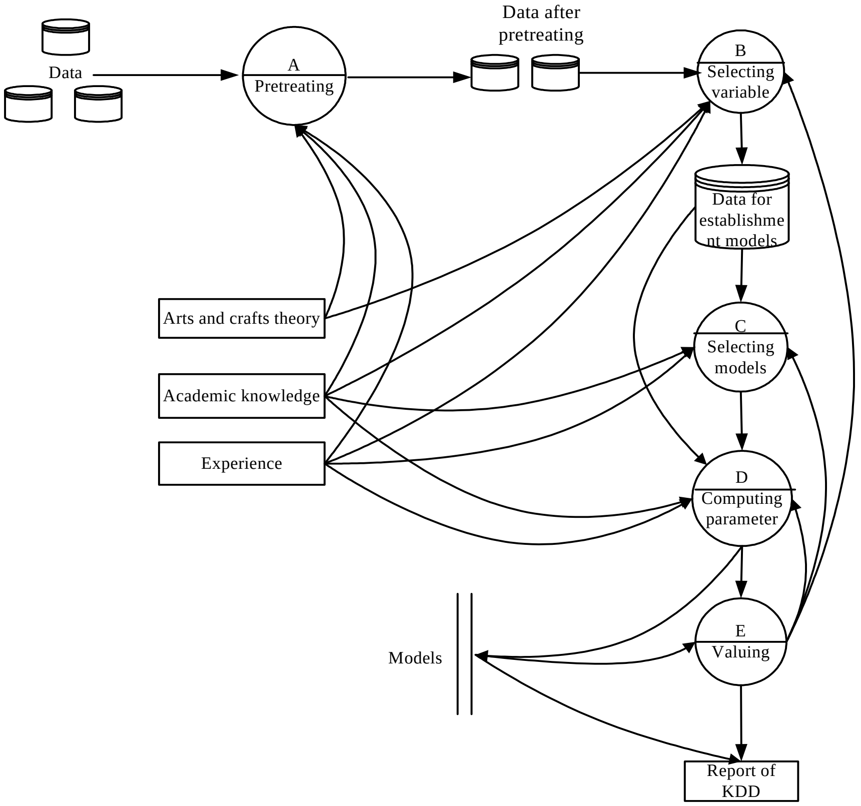

Fig. 1. The modeling process of data mining with embedded knowledge

Now there are some kinds of techniques applied in data mining, for example, the Artificial Neural Network(ANN), the decision tree, the genetic algorithm and the rule inferring, et al.. 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 methods and techniques of data mining becomes one of the most challenging problems 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 more analysis techniques applied in huge data information. Certainly, this task requests us to synthetically apply the relative knowledge of different disciplines. Based on these thoughts, we provide data mining techniques for time sequences based on Grey system theories in the next section.

3. Data Mining Techniques Based on Grey System Theories for Time Sequence Data

One of the main tasks facing the theories 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, et al [J.L.Deng,1985][S.F.Liu,1998]. In Grey system theories, it is through the organization of raw data for one to sort out development laws, if any. In other words, this is a path of finding out realistic governing laws from the available data. It is believed that even though objective systems phenomena can be complicated and related data chaotic, they always represent a whole, hence, implicitly contain some governing laws in theories of Grey system. The key for us to uncover and to make full use of all these laws is how to choose appropriate methods. The randomness of all grey sequence can be weakened to show its regularities through some generations. The operator theories provided by professor S.F.Liu, succeeds to solve the difficult problem of data pretreatment. The purpose of introducing buffer operators is to eliminate the shock waves or the noise that system behavioral data is interfered in order to show the true face of the data collected, based on conclusions of qualitative analysis.

In the view of data mining techniques with embedded knowledge, we think that modeling of Grey system itself is a kind of KDD and the data of economic phenomena are often regard as the time sequence data. Thus, seeing into thoughts of data mining with embedded knowledge and application the present Grey system theories, we can provide the data mining methods for time sequence data based on Grey system theories. The methods set is listed as follows.

Generation technique of grey sequence: to realize the data pretreatment with analysis of object system and applying scientific operators of sequences.

Grey incidence analysis techniques: to dig out the relationship quality among time sequences based on the geometry comparability of these sequences.

Grey incidence clusters analysis techniques: to class sub-groups based on the incidence analysis of data sequences and critical value.

Grey prediction techniques: to mine out the potential rule by the pretreatment of original data sequence, and interconvert grey difference equations to grey differential equation to establish a dynamic and continuous differential equation with discrete data sequences. Last, to realize the prediction of time sequences.

This paper only gives a demonstration of data mining techniques based on GM(1,1), which is the core of Grey system theories. Fig. 2 will show its modeling process.

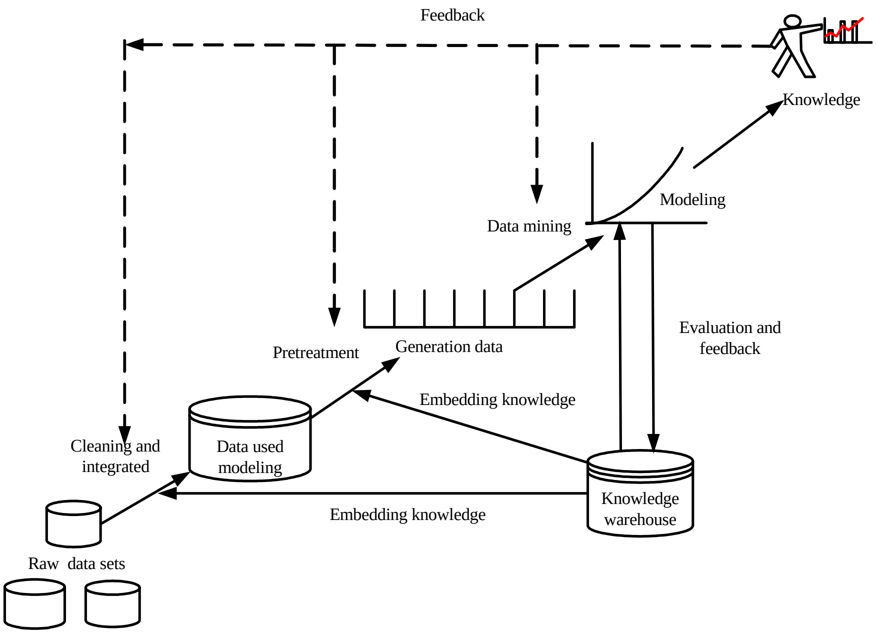

Fig. 2. The modeling process of data mining techniques based on GM (1,1)

Its idiographic arithmetic is described with pseudo-codes as follows.

```
Step 1: $p \leftarrow 0.98$    &&Initialize the average simulation precision $p$ required by the forecasting model.(Usually $p$ >98%)
Step 2: $X^0 \leftarrow (x^0(1), x^0(2), \cdots, x^0(n))$    &&Input the raw sequence.
Step 3: For $i \leftarrow 1$ to $n$
    $x^1(i) \leftarrow \sum_{k=1}^{i} x^0(k)$  && 1-AGO(Accumulating Generation Operators)
Next
Step 4: For $i \leftarrow 2$ to $n$
    $z^1(i) = x^1(i) + x^1(i-1)$    && Compute the generated mean value of the consecutive neighbors of $X^1$.
Next
```

```
Step 5: Find $a$, $b$.    && Base on the modeling steps of GM(1,1).
Step 6: For $i \leftarrow 2$ to $n$
    $\hat{x}^{(0)}(k) \leftarrow (x^{0}(1)-\frac{b}{a})(e^{-a}-1)e^{-a\cdot(k-2)}$
    && Compute the simulation value of all data.
    Next
Step 7: Find $\varepsilon$ and $p^{'}$.    && Compute the average relative simulation error $\varepsilon$ and the precision $p^{'}$.
Step 8: IF $p^{'} \geq p$ then
    For $i \leftarrow n+1$ to $n+L$    && Carry through predictions of $L$ steps.
    $\hat{x}^{(0)}(k) \leftarrow (x^{0}(1)-\frac{b}{a})(e^{-a}-1)e^{-a\cdot(k-2)}$
    Next
    Go to step 9
    Else
    $X^{0} \leftarrow X^{0}D$    && Inflict certain kind of buffer operator, which should blend the qualitative analysis into the system model.
    Go to step 3
    Endif
Step 9: Output $a, b$ and simulation values, simulation errors, average relative errors and the required prediction values, respectively.
Step 10: End
```

There are several kinds of weakening operators and strengthening operators in Grey system theories to weaken and to strengthen the increasing trend of time sequences, respectively. We can establish the scientific operator based on practices, and apply it to data pretreatment in the process of system analysis.

4. Demonstration

Give a demonstration with the total homes connecting with Internet (unit: Ten Thousand Homes)[Yearbook,2002] to show the prediction techniques based on GM (1,1).

According to the consistency and relativity of prediction and the “new information first” thoughts of Grey system, we select the total homes of 1996~2001 connecting with Internet as the original sequence.

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

We can find that its trend is rapid and there is nearly 1~10 times growth rate per year. Certainly, it cannot keep up with so high growth rate in the Chinese economic situation present, so, it isn't receivable and approbatory to establish prediction model directly with raw data. With careful analysis in depth, we think that the reasons of higher growth rate are due to the low baseline, and the low baseline is a consequence of the fact that in the past, the Internet just now is springing up and it needs a process to accept a new birth thing. Thus, we must weaken its growth rate and consider the process into the sequence when forecast the total home connecting with Internet of future years. A kind of weakening buffer operator is shown as follows.

Assume that the sequence of raw data and one of its buffer sequences are $X = (x(1), x(2), ..., x(n))$

and $$XD = (x(1)d, x(2)d, ..., x(n)d)$$

where $$x(k)d = \frac{1}{n-k+1}(x(k) + x(k+1) + ... + x(n))$$

for any $k = 1,2,......,n$, we call XD is the weakening buffer operator, which can weaken the growth trend of sequences.

Furthermore, we introduce the above weakening operator and then obtain the first order buffer sequence as follows

XD =(41.0085,49.144,61.205,78.193,96.17,104.10)

and introduce the second-order weakening operator and then obtain the following second-order buffer sequence

XD^2 =(71.637,77.762,84.917,92.821,100.315,104.10).

Based on the above idiographic arithmetic, we assume the required simulation precision is 98%, and then establish the forecasting models with the above arithmetic. The GM(1,1) model with non-pretreatment, with XD-pretreatment, and with XD^2-pretreatment is listed as follows.

$$\hat{x}(1996+t) = 26.582122e^{0.53093t} - 26.252122$$

$$\hat{x}(1996+t) = 261.56478e^{0.182115t} - 220.55628$$

$$\hat{x}(1996+t) = 1040.9309e^{0.073135t} - 969.2939$$

Then, with three models, we can get the simulation precisions $p^{'}$, the relative errors ε shown in table one as follows.

Table 1. Comparing among three models

Model	Data pretreatment	ε (%)	$p^{'}$ (%)	p (%)	Whether or not meet with the required precision
1	No	451.8	<0	98	No
2	XD	4.54	95.46	98	No
3	XD^2	1.31	98.69	98	Yes

From table 1, we can find only the prediction precision of the third model is bigger than the required one, and it is obvious that data prediction model with suitable pretreatment is benefit to improve the precision of simulation and prediction. Because it meets with our requirement to establish the model with the second-order buffer operator XD^2, we can get the suitable prediction model with XD^2.

$$\hat{x}(1996+t) = 1040.9309e^{0.073135t} - 969.2939$$

Its average simulation error is only: 1.31%, and the total homes connecting with Internet of 2002~2005 are gotten respectively: 113.851, 122.49, 131.784, 141.783.

5. Conclusion

From the above we know that:

This paper provides the data mining methods techniques for time sequence based on Grey system theories and the thoughts of data mining with embedded knowledge first. These methods will richen the data mining techniques, especially for time sequences. Certainly, all data mining techniques have its suitable range, and we should pay more attention to the analysis of systematic information and systematic phenomenon to adopt the scientific and appropriate pretreatment when we apply these data mining techniques based on Grey system.

Additionally, it is known that it is a complex and difficult question how to embed knowledge into the mathematic models with some model pattern when we apply the data mining techniques based on Grey system theories, but Prof. Qian Xuesen' s meta-synthesis [J.Y.yu,2001][J.Y.Yu,2002] can give us some helps and reveals to solve these questions. In future researches, we will succeed our attentions into data mining techniques for time sequence data based on Grey system theories with more meticulous modes, and pay more attentions to the generation technique of grey sequence, the grey incidence analysis techniques, the grey incidence cluster analysis techniques, which all be the component of data mining techniques based on Grey system theories. Furthermore, it is more important that we will apply these data mining techniques into more wide fields, such as customer relation management, decision-making under uncertainty, et al.

Acknowledgement

The authors thank the editor and two anonymous referees for their helpful comments and suggestions, which have improved the exposition of this paper. Additionally, this paper is partially supported by the Key Program of

Natural Science Foundation of China (BK2003211), Research Foundation for the Doctoral Program of National Ministry of Education of China (20020287001), and Nature and Science Foundation of Education committee of Henan Province(2006120004).

6. References

1. J.L.Deng(1985).Grey Systems: Society and Economics. Beijing: Press of National Defense Industry.
2. J.Y.Yu(2001).Qian xue-Sen's contemporary system of science and technology and meta-synthesis. Chinese engineering science.11:10~18.
3. J.Y.Yu and Y.J.Tu(2002). Meta-synthesis-Study of Case. System engineering theory and practice.5:1~7,42.
4. Jia-Wei Han, Kamber,M. Translation by Ming Fan(2001).data mining: conception and techniques.mechanic press.
5. S.F.Liu and Y.L(1998). introduction to Grey Systems: Foundations, Methodologies and Applications. Slippery Rock, IIGSS Academic Publisher.
6. S.F.Liu, T.B.Guo and Y.G.Dang.Grey(1999). Systems Theory and its applications. Science Press.
7. Stat. Yearbook of Shanghai.Stat. Press.2002.9
8. Y.J.Wang(2001).Practical modeling methods for SCM and data mining. Tsinghua University Press.

Bin Liu, received his B.S. degree in mathematics from Henan University, Kaifeng, China in 1994, and PhD in management science and engineering from Nanjing University of Aeronautics and Astronautics(NUAA), Nanjing, China in 2005. Now he is the post-doctor in control science and engineering at NUAA, and also the associate professor and director in department of management science, Henan Agricultural University. His main research interests include information system, supply-chain management, and grey system theories. Since 2000, he has published 28 technical papers in various journals, such as IEEE Transaction on System, Man and Cybernetics-Part A: systems and humans.

Hui Zhang, received her B.S. degree in computer science from Henan Normal University, Xinxiang, China in 1997, and M.S. degree in Software Engineering from Zhengzhou University, Zhengzhou, China in 2005. Now, she is with department of computer science, Henan Agricultural University, Zhengzhou, China. Her main research interests include information system, and data mining method.

Sifeng Liu, received his B.S. degree in mathematics from Henan University, Kaifeng, China in 1978, and the M.S. and PhD degrees, both in system engineering, from Huazhong University of Science and Technology, in 1986 and 1998, respectively. Now, he is a distinguished professor at NUAA, and also the academic leader in system engineering discipline. His main scientific research activities are in grey system and grey information modeling, and econometrics.

Yaoguo Dang, received his B.S. degree in mathematics from Henan University, Kaifeng, China in 1986, and the PhD degree in management science and engineering from NUAA in 2006. Now, he is with NUAA, and also the professor and director in department of management engineering. His main research interests are in grey system and grey information modeling, and forecasting methods.