

A Data Mining Approach Based on Grey Prediction Model in Web Environment

Hui Peng^{1,2}

*Hunan Knowledge Grid Lab¹, School of Computer Science and Engineering²,
Hunan University of Science and Technology, Xiang Tan, China, 411201
penghui1999@sohu.com*

Abstract

This paper proposes a prediction approach that combines grey prediction model with three-layer computing architecture in Web environment. It presents a refined prediction formula to enhance the scalability and usage scope of the Web prediction systems. It layouts a three layer architecture for the Web prediction system to reduce the network traffic and maintain load balance. A traffic decision support system is presented to illustrate the proposed approach.

1. Introduction

Prediction plays an important role in data mining. Grey prediction model applies consecutive differential equation to perform a general analysis and a long-term prediction based on a small quantity of discrete data [1, 2]. Since grey prediction model can deal with insufficient information and half-baked data, it is usually employed in information processing systems in fields such as sociology, ecology and economics, where the data collected from these fields are usually discrete, inadequate and irregular.

However, most these systems are usually committed to isolated local prediction system, which limits their application scope and information sharing. A general approach to relieve this suffering is to extend these systems that apply the grey prediction model to Web applications[5, 6], so that more people can use these information resources via the World Wide Web to support their decision process[7,9]. In this case, two crucial problems should be solved: (1) encapsulate an algorithm and keep it isolated from other parts of a Web system, and (2) appropriately distribute computing tasks of an application to different processing nodes to reduce the network traffic and maintain load balance. To solve the first problem, a

concise prediction formula was refined after roundly analyzing the grey predict algorithm. This enhances the scalability and usage scope of the Web prediction systems. A three layer architecture for the Web prediction system was proposed to resolve the second problem. Moreover, an additional original error parameter is also added to the primitive prediction model to improve predict precision. Finally, a traffic decision support application is presented to illustrate the proposed approach.

2. Grey Prediction Model

2.1. Grey Prediction Algorithm

The grey model GM(1,1) and the grey modeling procedure of GM(1,1) [1,2] are briefly described as follows:

$x^{(0)}(0), x^{(0)}(1), \dots, x^{(0)}(n-1)$ is a non-negative original data sequence, let $x^{(0)}$ be a stochastic variable restricted by $x^{(0)}(0), x^{(0)}(1), \dots, x^{(0)}(n-1)$, take the accumulated generating operation on $x^{(0)}$, we can obtain the sum data sequence $x^{(1)}(0), x^{(1)}(1), \dots, x^{(1)}(n-1)$,

$$x^{(1)}(k) = \sum_{m=0}^k x^{(0)}(m) \quad k = 0, 1, \dots, n-1 \quad (1)$$

Let $x^{(1)}$ be a stochastic variable restricted by $x^{(1)}(0), x^{(1)}(1), \dots, x^{(1)}(n-1)$, the first order differential equation of $x^{(1)}$ is

$$dx^{(1)} / dt + ax^{(1)} = u \quad (2)$$

In order to find out the solution of the above differential equation, the parameters a and u must be decided. They can be solved by means of the least-square method as follows:

$$\hat{a} = \begin{bmatrix} a \\ u \end{bmatrix} = [B^T B]^{-1} B^T Y_N$$

and

$$B = \begin{bmatrix} -1/2[x^{(1)}(0) + x^{(1)}(1)] & 1 \\ -1/2[x^{(1)}(1) + x^{(1)}(2)] & 1 \\ \dots & \dots \\ -1/2[x^{(1)}(n-3) + x^{(1)}(n-2)] & 1 \\ -1/2[x^{(1)}(n-2) + x^{(1)}(n-1)] & 1 \end{bmatrix}$$

$$Y_N = [x^{(0)}(1)x^{(0)}(2)\dots x^{(0)}(n-1)]^T$$

By the theory of the first order linear differential equation, the solution of the equation (2) is $x^{(1)}(t) = (x^{(1)}(0) - u/a) \cdot e^{-at} + u/a$, Replacing t with discrete subscription k , we can get the discrete solution of the equation

$$x^{(1)}(k) = (x^{(1)}(0) - u/a) \cdot e^{-ak} + u/a \quad (3)$$

From (3), let $k = 0, 1, \dots, n-1$, we can get the predict data sequence $\hat{x}^{(1)}(0), \hat{x}^{(1)}(1), \dots, \hat{x}^{(1)}(n-1)$, it is the predict data sequence of $x^{(1)}(0), x^{(1)}(1), \dots, x^{(1)}(n-1)$,

So

$$\hat{x}^{(1)}(k) = (x^{(1)}(0) - u/a) \cdot e^{-ak} + u/a \quad (4)$$

because

$$\hat{x}^{(1)}(k) = \sum_{m=0}^k \hat{x}^{(0)}(m), \quad k = 0, 1, \dots, n-1, n, \dots \quad (5)$$

from (4),(5), there exists

$$\begin{cases} \hat{x}^{(0)}(k) = [x^{(0)}(0) - u/a](1 - e^{-a})e^{-ak} & k=1, 2, \dots, n-1, n, \dots \\ \hat{x}^{(0)}(0) = x^{(0)}(0) & k=0 \end{cases} \quad (6)$$

It is the predict formula for the original data sequence $x^{(0)}(0), x^{(0)}(1), \dots, x^{(0)}(n-1)$ ($n \geq 3$), let $k = n$, we get the first predict value, $k = n+1$, the second, and so on. But before actual prediction stars, the prediction model must be adjusted to fit the defined error precision. The method of adjusting model is described in 2.2.

2.2. Adjusting Model

The variance ratio checking method is mostly used to check the precision of the grey model. The two parameters the method uses to adjust the model is the variance ratio c and the minor error frequency p ,

Let $\varepsilon^{(0)}(k) = x^{(0)}(k) - \hat{x}^{(0)}(k)$ ($k = 0, 1, \dots, n-1$),

$$\bar{\varepsilon} = 1/n \cdot \sum_{k=0}^{n-1} \varepsilon^{(0)}(k), \quad \bar{x} = 1/n \cdot \sum_{k=0}^{n-1} x^{(0)}(k)$$

$$S_1^2 = 1/n \cdot \sum_{k=0}^{n-1} (\varepsilon^{(0)}(k) - \bar{\varepsilon})^2$$

$$S_2^2 = 1/n \cdot \sum_{k=0}^{n-1} (x^{(0)}(k) - \bar{x})^2$$

$$c = S1/S2. \quad (7)$$

$$p = P\left\{ \varepsilon^{(0)}(k) - \bar{\varepsilon} < 0.6745 * S_2 \right\} \quad (8)$$

From (7), (8), we can get the grade of a prediction, if the precision is accord with practical request, we can apply the model to execute actual predict. Otherwise we should use error sequence $\varepsilon^{(0)}(k)$ ($k = 0, 1, \dots, n-1$) to adjust the predict model. The adjusting method is: use error sequence as original data sequence, apply the computer procedure in 2.1 to the sequence, we can get $\hat{\varepsilon}^{(0)}(k)$ ($k = 0, 1, \dots, n-1$), add

$\hat{\varepsilon}^{(0)}(k)$ with $x^{(0)}(k)$, we can get the next generation of original data sequence, apply the computer sequence in 2.1 to the new sequence, we can get the next generation of predict data. According to the above description, the following formula can be used to get the new predict data:

$$\begin{cases} \hat{x}^{(0)}(k) = [x^{(0)}(0) - u_0/a_0](1 - e^{-a_0})e^{-ak} + [\varepsilon^{(0)}(0) - u_1/a_1](1 - e^{-a_1})e^{-ak} & k=1, 2, \dots, n-1, n, \dots \\ \hat{x}^{(0)}(0) = x^{(0)}(0) & k=0 \end{cases}$$

then apply the adjust procedure to the new sequence,

when the next generation of $\hat{\varepsilon}^{(0)}(k), c, p$ is generated, if the predict precision is satisfied, we can apply the model to execute actual predict. Otherwise add the new generation $\hat{\varepsilon}^{(0)}(k)$ to the original data sequence

$x^{(0)}(k)$ to start the next generation of predict. Repeat the above procedure until final results meet satisfaction or procedure terminates on the defined times of repeat which means the original data can not be predicted.

3. A Data Mining Approach in Web Environment

3.1. The Structure of Web Application System

Traditional Web application system structure is server/client structure. One method of the structure is server works as the core of a system and is responsible for all data processing and client only display HTML pages. In this method, the server's load is heavy, and all results return to client increase the flux of network. The other method of the structure is "fat" client that means data process concentrates on client. This method reduces the load of server and the flux of network but heavy the load of client, reduces the speed of application system and makes maintenance of system inconvenient. The three layer structure of Web application system includes three layers: expression layer, logic layer and data layer. Expression layer is the user's interface. It displays the data and communicates

with logic layer: It accepts inputs (the request of data processing) from user, passes the request to logic layer, accepts the results of data processing from logic layer and display the result reasonably. Logic layer is the core of a system. It processes data and communicates with expression layer and data layer: It accepts the request from expression layer, processes data which fetched from data layer according to request and returns results to expression layer. Data layer saves data. It often includes a database server and some data files.

The three layer structure solves the problems which the above two methods produce. It distributes the calculation to both server and client, balances the load and flux of network.

3.2. The Structure of Web Application System

We propose a data mining approach which combines grey prediction model with three layer structure in Web environment. Figure 1 describes the structure of a system which applies this approach.

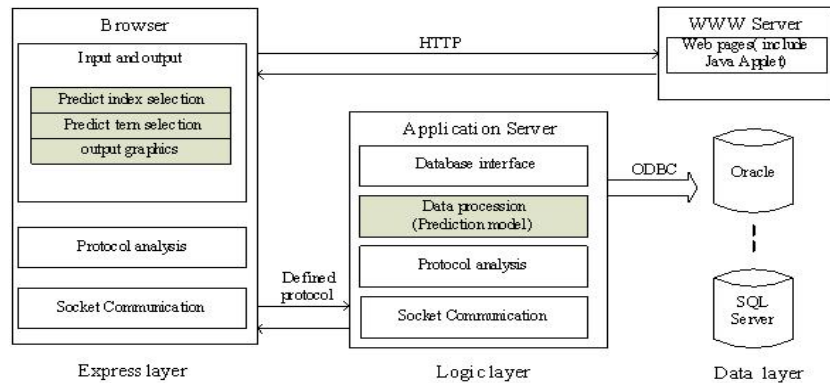


Figure 1. The Structure of a System Which Applies Grey Prediction Model

Start the application server first when the system runs. The server will listen to the connection from client. The client downloads HTML pages include Java Applet from server through HTTP protocol and display them in browser. The connection between server and client is built when Java Applet runs[3,4]. After that, server and client communicate through self-defined protocols. Applet accepts prediction parameters and passes them to application server. Application server fetches data from database according to parameters from client, applies grey prediction model to process data and returns prediction results to Applet. Applet displays prediction results and draws prediction trend figure. All original data is located in database server. The design of application server is the key of the system. It mainly includes the implement of grey prediction model and the self-defined protocol

(2) It balances the load on Internet. The special application server process data makes “fat” client to be “slim”, the data processing ability of client reduces the load of server also. It also controls the information flux in internet and avoids the network block.

(3) The installation and maintenance of system software is facility. No client installation in this method. All system maintenance locates in server. So it is easy to update the system software.

(4) The structure can be easily succeeded by other web application system. Replacing the “prediction model” with other models in “data procession” part and replacing the “input and output” part in the Figure 1(the grey part in figure 1) will produce other web applications of this method.

3.3. The advantages of this method

The advantages of the method lie in the following four respects[8,10]:

(1) The structure makes full use of the computing ability of both server and client. Data process is not concentrated on server or client any more and is distributed on both server and client.

4. An Application of the Data Mining Method

4.1. Traffic Decision Support System

The traffic department accumulated amounts of statistics data in database. To predict the traffic development trend from these data is very important and meaningful[4]. We apply the above data mining method in a traffic decision support system. 50 traffic

indexes are predicted. There are 95% predict value is above “good” precision according to experienced value. Figure 2 is the results of self check. The real prediction can start only when the precision grade of self check is satisfied. The “-8” in the figure means that building prediction model with 8 years data before nowadays. Figure 3 is the results of prediction of some index. The “+8” in the figure means that to predict 8 years data after nowadays. The traffic indexes and the predict years can be adjusted. There are 50 traffic indexes altogether and the predict years change from -8 to 15.

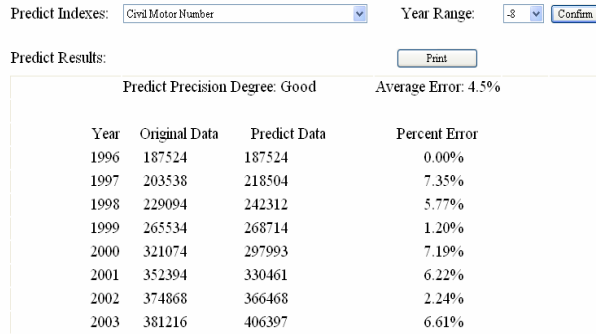


Figure 2. The Results of Self Check with the Traffic Index of Civil Motor Number

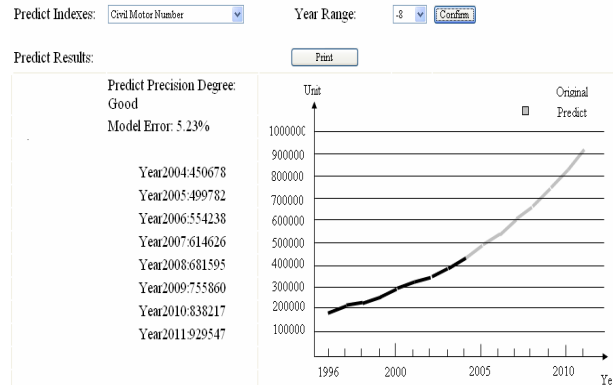


Figure 3. The Predict Results of the Traffic Index of Civil Motor Number

4.2. Several Notable Problems in Programming

(1) The grey prediction model may be found with three original data. But in application system we usually use five to eight original data. It can improve prediction precision. In the above traffic application system we use eight original data to start a prediction.

(2) To improve the precision of prediction, we can add some new error parameters to decide the precision of a prediction in the procedure of model adjustment. In the above traffic system, we use origin error parameter \bar{e} with c and p to decide the precision of

predict. $\bar{e} = 1/(n-1) \cdot \sum_{k=2}^n |\varepsilon^{(0)}(k)| * 100\%$.

(3) The effect of model adjustment will decline as the increase of adjust times. We have to balance the precision and the computer time in application system. In the above traffic application system we defined the maximum times of model adjustment is 100.

(4) The original data of the grey predict algorithm should be positive. If there exists negative data in error sequence, a positive offset should be added to the error sequence, and the offset should be subtracted from the predict results.

5. Conclusion

A suitable predict algorithm and an appropriate structure is crucial for the Web prediction system. Our proposed approach that combines the grey prediction algorithm with three layer structure enhances the scalability and usage scope of the grey prediction model. It was applied to a traffic decision support system. The system represents the predict results and the development trend graphics for each traffic indexes. 50 traffic indexes are predicted successfully.

6. References

- [1] J. Deng. The Basic Method of Grey Theory. Huazhong University of Science and Technology Press. (1987)
- [2] J. Deng. The Grey Control System. Huazhong University of Science and Technology Press. (1993)
- [3] J. Jaworski. Java Developer's Guide. Sams.net Press (1996, in Chinese)
- [4] Tianyue, the design and development of network traffic information system, Hunan traffic technology.6(1995, in Chinese)
- [5] H.Zhuge, The Knowledge Grid, World Scientific Publishing Co., Singapore, 2004.
- [6] H.Zhuge, Discovery of Knowledge Flow in Science, Communications of the ACM, 49 (5) (2006) 101-107.
- [7] H.Zhuge, The Future Interconnection Environment, IEEE Computer, 38 (4) (2005) 27-33.
- [8] H.Zhuge, et al, A Scalable P2P Platform for the Knowledge Grid, IEEE Transactions on Knowledge and Data Engineering, 17 (12) (2005) 1721-1736.
- [9] H.Zhuge, China's E-Science Knowledge Grid Environment, IEEE Intelligent Systems, 19 (1) (2004) 13-17.
- [10] H.Zhuge, Resource Space Grid: Model, Method and Platform, Concurrency and Computation: Practice and Experience, 16 (14) (2004) 1385-1413.