The Performance Prediction of Cloud Service via JOGM(1,1) Model

Zhipiao Liu, Qibo Sun, Shangguang Wang, Fangchun Yang

Abstract

With the rapid development of cloud computing, more and more cloud services are provided to users. Faced with multiple cloud services, how to scientifically predict and assess the performance of cloud service is an imperative task. To cope with the challenge, a prediction scheme via JOGM(1,1) model is proposed based on grey system theory, in which the performance of cloud service is quantified as the response time. Thus cloud users and the third-party institutes of cloud service performance evaluation can predict and assess the performance of cloud service as accurate as possible. In return, this will contribute to the cloud service provider selection for better performance on behalf of cloud consumers. The simulation results show that the proposed scheme has higher prediction precision compared with classic GM(1,1) model and weighted moving average model, which verifies the effectiveness of the proposed prediction scheme and the feasibility of forecasting the performance of cloud service employing grey system theory.

Keywords: Cloud Service; Performance Prediction; Response Time; JOGM(1,1); EC2

1. Introduction

As a promising computing paradigm, cloud computing has drawn extensive attention from academia and industry in recent years. Cloud computing services allow users to lease computing resources from large scale data centers operated by service providers. Using cloud services, users can deploy a wide variety of applications dynamically and on-demand [1].

Along with this new paradigm, cloud computing can be divided into the following three service models: IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service) [2]. Nowadays almost every well-known IT companies such as Amazon, Google, IBM and Salesforce, have introduced related cloud services. Amazon’s EC2/S3, Google’s App Engine and Salesforce’s online CRM are successful commercial examples.

On the one hand, compared with traditional desktop computing, cloud computing presents many advantages, such as flexible scalability, pay-as-you-use, high resource utilization and economies of scale. On the other hand, with the advancement of cloud computing technologies including SOA, virtualization, security and high bandwidth network access, large numbers of business applications from companies and institutes will be moved into the cloud in the near future [3, 4, 5, 6]. In order to grasp the emerging business opportunities, different cloud service providers are racing to deliver their own cloud services. However, many cloud services derived from different cloud providers are homogeneous. In other words, these services have equivalent functions, but different performance. Faced with these homogeneous cloud services, how to scientifically assess their performance has become a practical requirement.

Cloud service performance is concerned with how fast the cloud can provide the requested service, which is usually quantified by the service response time, i.e., the time elapsed from a user’s submitting a service request to the cloud until the user receives the final responsive results from the cloud [7, 8]. Long service response time means low performance. It is evident that if the cloud service performance is low, such as a poor cloud storage service [5], the cloud users will frequently suffer long storage service response time, which could lead to users’ fret and dissatisfaction. As a result, those impatient users are likely to resort to other cloud providers, perhaps your competitors, even though it has low...
reliability but high performance. Contrary to this situation, if your cloud service is responsive, then you will witness increasing market share and more profits.

With respect to the significance of cloud service performance, the accurate prediction of cloud service response time is especially important both for cloud providers and cloud users. In fact, the response time of cloud service relies heavily on the ability of cloud service delivery platform and network stability [9]. Typically, the cloud consumers could not acquire the internal details of cloud services provision platform in a high dynamic and stochastic network environment. Therefore, as an essential means of service evaluation, the design of performance prediction scheme is particularly important for scientific assessment of cloud services.

To deal with this practical challenge and improve the prediction accuracy of the response time of cloud service, we conduct detailed research on traditional GM(1,1) prediction model, and propose an new joint optimization grey model named JOGM (1,1), which employs the recent sequence of monitored response time as input data to predict the response time of cloud service in the near future based on poor information and insufficient samples. Experimental results show that our proposed prediction approach obtains better prediction accuracy than traditional GM(1,1). The obtained prediction data can help decision makers in companies or institutes make informed choices for cloud service with best performance.

The rest of this paper is organized as follows: Section 2 introduces background and related work. Section 3 presents our JOGM(1,1) prediction model, and then a time series prediction scheme based on JOGM(1,1) is proposed. In section 4, we illustrate that our proposed scheme has better prediction performance than traditional GM(1,1) model in terms of its prediction accuracy and validated the feasibility of the forecast scheme. Section 5 concludes the paper.

2. Background and related work

For one thing, faced with multiple cloud services, cloud users, whether start-up or personal, have neither enough time nor adequate budget to test all cloud services one by one. For another, a large number of cloud service providers continue to publish new cloud services to the Internet for users. As a result, it maybe lead to such an embarrassing situation in which cloud users have little understanding of the performance of candidate services, probably only a few groups of historical performance data.

Therefore, under the condition that cloud user usually has very limited performance data on cloud services, how to forecast accurately the performance of cloud services of the next steps based on “incomplete information and small sample”, and allowing cloud users to choose the best cloud service providers with the help of prediction results, now becomes an impending problem to be solved.

A thorough literature survey reveals that while there has been a lot of work in the area of assessing the performance of virtual resource or cloud service [8, 9, 10], very little attention has paid to the performance prediction of cloud service in term of cloud service provider selection.

Simson L. [11] makes an end-to-end performance analysis of Amazon S3’s throughput and latency as observed from Amazon’s EC2 cluster and other locations on the Internet; and an analysis of the Amazon SQS operation and performance, but do not refer to the problem of performance prediction.

Close to our work is the seminal study of the relationship between the change to infrastructure deployment and the service response time [12], which presents What-If Scenario Evaluator (WISE), a tool that can quickly and accurately predict service response-time distributions for many practical what-if scenarios. On the basis of the above study, Wang et al. [7] not only provides a tool to measure service performance accurately, but also propose new methodologies for cloud service providers to evaluate hypothetical new deployments. However, the above-mentioned two related studies are all carried out from the perspective of service providers in order to optimize the configuration and deployment of infrastructure, which is different from our work derived from the perspective of cloud service users.

Similarly to our work, from the cloud service user’s point of view, Iosup et al. [13] analyze the performance variability of cloud services from Amazon Web Services and Google App Engine, currently two of the largest commercial clouds in production, and assess the impact of the variability observed for the studied cloud services on three typical large-scale applications in terms of cloud provider selection. The performance study of cloud service [13], however, focuses on the performance variability and evolution of cloud service. Our work complements this study by concentrating on the
average performance values over time, especially the responsive time prediction value in the near future, which will give evidence that predictive performance value is an important factor in cloud service provider selection. In other words, without internal information of cloud service provision platform and influenced by the dynamic network conditions, we propose a performance prediction approach to forecast the response time of cloud service as accurate as possible, only based on the limited recent sample data, without requiring real-world cloud service invocations.

3. The prediction scheme via JOGM(1,1)

Being able to forecast time series accurately has been quite a popular subject for researchers both in the past and at present [14]. Taking into account the high complexity of cloud service infrastructure and the randomness of network conditions, the whole cloud provisioning system can be seen as a grey system to study. Grey System Theory (abbreviated as GST) was first proposed in early 1982 by Deng to study uncertainty of system with small amount of data and incomplete information [15]. Derived from the conception of black box and white box, a system with partial information known and partial information unknown is grey system. The theory is composed of five parts, which includes grey prediction, grey relation, grey decision, grey programming and grey control. In recent years, the methodologies of grey prediction based on grey theory have been successfully used in many fields and demonstrated promising results.

As a time series forecasting model based on GST, GM(1,1) which stands for the first order grey model with a single variable, usually requires only 4–10 recent sample data to predict the future values of a time series. Generally speaking, a prediction model requires $n$ previous value as input, and produces $m$ predicted value as output. The input data are called samples, and $n$ is the sample number. According to statistical theory, if $n < 30$, the sample is small [16]. Compared to traditional statistical prediction approaches, such as neural networks or fuzzy models, GM(1,1) model eliminates the inherent defects requiring large sample to form the prediction model. Therefore, in terms of the performance prediction of cloud service, GM(1,1) is very applicable, especially when only very limited sample data can be obtained by cloud users.

3.1. JOGM(1,1) model

As the basic form of grey model, classic GM(1,1) model has been widely applied to address real-world problems. However, there is still much room to improve in terms of the prediction accuracy of GM(1,1). In fact, grey model prediction can be viewed as a curve fitting approach [17], whose prediction precision depends on the approximation degree of exponential curve fitting, and the approximation degree relies on the parameters of grey model. Therefore, any effective and efficient optimization method for GM(1,1) model should take this key point into consideration.

3.1.1. The optimization of the initial value

When it comes to GM(1,1), we try to use the exponential equation

$$x^{(i)}(k + 1) = C e^{a(k + 1)} + b, \quad k = 1, 2, \ldots, n - 1. \quad (1)$$

to fit the pre-processed data sequence $x^{(i)}(k)$ via Least Square Method. However, inspired by the theory of Least Square Method, we know that the above fitting curve obtained from the whitening equation $dx^{(i)} / dt + ax^{(i)} = b$, may not pass through any point of the set $\{(m, x^{(i)}(m)) | m = 1, 2, \ldots, n\}$. But we have to assume that the above fitting curve pass through one of these points in order to get the concrete value of coefficient $C$. In original GM(1,1) model, we always suppose that the fitting curve pass through the point $(1, x^{(0)}(1))$, and then get the coefficient value $C = x^{(0)}(1) - b / a$.

However, the above assumption violates the important principle of new information priority in grey model [15], which means that GM(1,1) may have better performance in predicting time series only based on the latest data instead of using the relatively old $x^{(0)}(1)$ to fit curve. In fact, we can get the more general solution, as shown below.

We let
\begin{align*}
\hat{x}^{(i)}(m) &= x^{(i)}(m) = Ce^{-at(m-n)} + b/a, \quad m = 1, 2, \ldots, n. 
\end{align*}

In other words, \( x^{(i)}(m) \) is used to be the initial condition to fit, and then figure out the coefficient value, that is
\begin{align*}
C &= (x^{(i)}(m) - b/a)e^{at(m-1)}. 
\end{align*}

Thus we get the general solution expressed as:
\begin{align*}
\hat{x}^{(i)}(k+1) &= (x^{(i)}(m) - b/a)e^{at(m-k-1)} + b/a, \quad k = 1, 2, \ldots, n-1; \quad m = 1, 2, \ldots, n.
\end{align*}

Finally, we obtain the predicted value as follows,
\begin{align*}
\hat{x}^{(0)}(k+1) &= (x^{(i)}(m) - b/a)(e^{at(m-k-1)} - e^{at(m-1)}), \quad k = 1, 2, \ldots, n-1; \quad m = 1, 2, \ldots, n,
\end{align*}
and \( \hat{x}^{(1)}(1) = \hat{x}^{(0)}(1) = x^{(0)}(0) \), where \( \hat{x}^{(0)}(k+1) \) is the predicted value of \( x^{(0)}(k+1) \).

3.1.2. The optimization of the background value

As mentioned above, the prediction accuracy relies on the parameters of grey model. With this in mind, Luo Dang et al. [18] make a thorough examination of the background value \( z^{(i)}(k) \), on which the values of the development coefficient \( a \) and the grey action quantity \( b \) partly depends, and find that the traditional calculation formula \( z^{(i)}(k) = 1/(2(x^{(i)}(k) + x^{(i)}(k-1)), \quad k = 2, 3, \ldots, n \) is one of the key factors leading to the following simulation error in original GM(1,1), that is
\begin{align*}
\hat{e}^{(0)}(k) &= \hat{x}^{(0)}(k) - x^{(i)}(k)
\end{align*}

To eliminate the above error derived from \( z^{(i)}(k) \) calculated by the mean value of \( x^{(i)}(k) \) and \( x^{(i)}(k-1) \), we let
\begin{align*}
\hat{z}^{(i)}(k) &= \int_{t=0}^{k} x^{(i)}(t) dt
\end{align*}
and finally figure out the solution of the above integral equation as follows:
\begin{align*}
\hat{z}^{(i)}(k) &= \frac{x^{(i)}(k) - x^{(i)}(k-1)}{\ln x^{(i)}(k) - \ln x^{(i)}(k-1)}
\end{align*}

To enhance the prediction accuracy, we propose a joint optimization GM(1,1) model, which integrates the two optimization ideas of initial value and background value discussed above. The newly generated model is defined as JOGM(1,1).

3.2. The prediction scheme based on JOGM(1,1)

On the basis of JOGM(1,1) model, a proposed prediction schematic diagram is shown in Figure 1.

![Figure 1. The prediction scheme based on JOGM(1,1) model](image-url)
Here is the algorithmic process of the prediction scheme.

Step 1: Initialization of raw response time data series to ensure that each element of the sequence is positive. The initialized time series \( x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(n)\}, \quad n \geq 4 \) (9) is used as the input of the prediction algorithm, and \( n \) is the number of input sample data. The task is to predict \( x^{(0)}(n+p) \) where \( p \) is the forecasting step, \( p \geq 1 \).

Step 2: Applying Accumulated Generating Operator (abbreviated as AGO) to the initialized series of response time \( x^{(0)} \) to create the increasingly monotonous exponential sequence \( x^{(1)} = \{x^{(1)}(1), x^{(1)}(2), ..., x^{(1)}(n)\} \), (10)

where
\[
x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \quad k = 1, 2, ..., n.
\] (11)

Step 3: Get the optimized background value sequence \( z^{(1)} = \{z^{(1)}(2), z^{(1)}(3), ..., z^{(1)}(n)\} \), (12)

where
\[
z^{(1)}(k) = (x^{(1)}(k) - x^{(1)}(k-1)) / (\ln x^{(1)}(k) - \ln x^{(1)}(k-1)).
\] (13)

Step 4: Generating the following GM(1,1) model
\[
x^{(1)}(k) + a z^{(1)}(k) = b,
\] (14)

and its whitening equation
\[
dx^{(1)} / dt + ax^{(1)} = b,
\] (15)

where \( a \) is development coefficient, \( b \) is grey action quantity. Assuming that
\[
B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix},
\] (16)

we can get the parameter estimators by means of Least Square Method, that is
\[
\hat{u} = (a, b)^T = (B^T B)^{-1} B^T Y
\] (17)

Step 5: Optimizing the initial fitting condition to obtain the time response sequence Eq. 4 and the prediction formula Eq. 5 of JOGM(1,1), respectively.

Step 6: Determine the parameter \( m \) in Eq. 5, and get the instantiated prediction formula. Let \( m = 1, 2, ..., n \) to traverse every input sample data to figure out the most appropriate value of parameter \( m \) in Eq. 5, which minimizing the Mean Square Error (abbreviated as MSE) expressed as follow:
\[
\sum_{k=m}^{n}(x^{(0)}(k) - x^{(0)}(k))^2.
\] (18)

Step 7: According to the above instantiated prediction formula Eq. 5, the predicted value can be figured out.

Step 8: end.

4. Experiment

4.1. Experiment setup

To validate the effectiveness of our prediction scheme, we compare proposed JOGM(1,1) model with the classic GM(1,1) model and weighted moving average model (abbreviated as WMAM). Our primary metric for evaluation is prediction accuracy. Three evaluating criteria, namely the mean absolute error (MAE), mean relative error (MRE) and mean square error (MSE), are used to measure the prediction accuracy, which are calculated respectively as
The Performance Prediction of Cloud Service via JOGM(1,1) Model
Zhipiao Liu, Qibo Sun, Shangguang Wang, Fangchun Yang

\[ MAE = \frac{1}{n} \sum_{k=1}^{n} |x^{(0)}(k) - \hat{x}^{(0)}(k)| \]  
(19)

\[ MRE = \frac{1}{n} \sum_{k=1}^{n} \left| \frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)} \right| \]  
(20)

\[ MSE = \frac{1}{n} \sum_{k=1}^{n} (x^{(0)}(k) - \hat{x}^{(0)}(k))^2 \]  
(21)

where \( x^{(0)}(k) \) denotes the original data and \( \hat{x}^{(0)}(k) \) is the predicted value, and \( n \) is the number of predicted values. Three criteria are calculated by the data of final prediction output and original input.

The data used for experiment are from Cloudsleuth [19], which is a well-known cloud service performance monitoring website. In order to measure the performance of cloud service, the website operators developed a reference application, which is complex enough to be somewhat representative of an actual e-commerce application, thereby avoiding the perils of micro-benchmarks, but simple enough to be readily understood and easily configured. The first page of the reference application contains small static images interspersed with text. The second page contains a single large (approximately 1MByte) image, an advertisement supplied by an ad service, and a map provided by a map service. Analytics are embedded on both pages [19]. And then the reference application is deployed to Amazon Elastic Compute Cloud (EC2) [20]. Amazon is the most famous cloud service provider, and hosts a few popular cloud services, such as EC2 which provide virtualized on-demand computing resources on a pay-per-use model. Cloudsleuth focuses on monitoring the performance of Amazon Elastic Compute Cloud (EC2) and delivers the user experience chart and the consistency chart of EC2. The service response time of the reference application deployed on EC2 for past days has been collected by Cloudsleuth.

In terms of the optimal number of data points to establish the GM(1,1) model, it is still an unsettled question. GM(1,1) is a small sample prediction model, and the small sample usually has more accuracy than the large sample when establishing GM(1,1) model in theory [16]. Generally speaking, at least four data sets should be measured in advance to establish the GM.

Taking into account the context of data starvation discussed above, the sequence of raw data collected by cloud users has only very limited data. Therefore, in this study, the recent five response time values are used to form the JOGM(1,1) prediction model, and the next value is predicted, which can be expressed as follow,

\[ t_{n+1} = f_{JOGM}(t_n, t_{n-1}, t_{n-2}, t_{n-3}, t_{n-4}), \]  
(22)

where \( t_{n+1} \) is the predicted value of next unit time step and \( t_{n-i}, \ i = 0,1,\ldots,4 \) is the recent input sample data.

4.2. Experiment results and analysis

We present one real cases of prediction to validate the accuracy of our proposed model based on the above error metrics. The response time data of our reference application deployed on Amazon EC2 for past 15 unit time steps are adopted, which are derived from the users located in Washington. In order to compare the prediction results of the proposed prediction model with conventional model, we present the widely used WMAM to do the same experiment. In the context of data starvation, cloud users have very limited sample data supporting the formation of prediction model. The successes of some well-known time series forecasting models, such as autoregressive integrated moving average (abbreviated as ARIMA), neural network models, fuzzy system models and, rely on a law for the distribution of original series as well as a large amount of observed data [17]. Therefore, they are not suitable in our scenario.

The predicted results based on recent five-data modeling are plotted in Figure 2, from which we could safely conclude that the prediction schemes based on both GM(1,1) and JOGM(1,1) model can largely match the cloud dynamics. In contrast, the prediction curve of WMAM model cannot keep track of the original one. Note that we compressed and expanded part of the figure so that the results can be easy to see.
Figure 2. The comparison of predicted results

Compared with the original GM(1,1), our JOGM(1,1) has higher accuracy, although the predicted error always exists. The comparative analysis of prediction errors are listed in table 1 (smaller value means better performance).

Table 1. Comparison of three error criteria

<table>
<thead>
<tr>
<th>Cloud service</th>
<th>Error criterion</th>
<th>WMAM</th>
<th>GM(1,1)</th>
<th>JOGM(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon EC2,</td>
<td>MRE</td>
<td>0.09688</td>
<td>0.05591</td>
<td>0.04162</td>
</tr>
<tr>
<td>Washington</td>
<td>MAE</td>
<td>0.43710</td>
<td>0.26306</td>
<td>0.1996</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.25145</td>
<td>0.11282</td>
<td>0.08805</td>
</tr>
</tbody>
</table>

As is shown in the Fig 2 and Table 1, the proposed JOGM(1,1) improves the prediction accuracy and has smaller error, which provides an excellent approach to forecast the dynamic performance of cloud service. On the one hand, without internal information of the cloud service provision platform, without sufficient past response time data, and influenced by the unpredictable network conditions, the cloud service provision platform can be seen a grey system from a cloud user’s point of view. In this context, it is much more difficult to make accurate prediction using conventional forecasting models. As is shown in figure 2, the prediction result of WMAM is unsatisfactory due to its large error. On the contrary, GM(1,1) require only a limited amount of data to estimate the behavior of unknown systems, and thus largely match the cloud service’s dynamics. On the other hand, our proposed JOGM(1,1) model based on the original GM(1,1) further enhances the prediction accuracy. The predicted results validated the effectiveness and feasibility of proposed prediction scheme. Therefore, it is very applicable to forecast the highly dynamic performance of cloud service using grey system theory.

5. Summary

Cloud computing are becoming new platforms for enterprise and personal computing. The next years will witness more and more cloud services on the Internet. In order to scientifically predict and assess the performance of cloud service, we propose an optimized JOGM(1,1) model, and then present an prediction scheme based on the proposed model, which can be used to forecast the performance of cloud service with limited latest sample data. The simulation results show that the proposed scheme has higher prediction accuracy compared with classic GM(1,1) model and WMAM model, which verifies the effectiveness of the proposed prediction scheme, and the feasibility of forecasting the highly dynamic performance of cloud services with the help of grey system theory.

6. Acknowledgement

We would like to thank the reviewers and the relative editors for their constructive comments and kind consideration. Our work is supported in part by the Specialized Research Fund for the Doctoral
Program of Higher Education under grants no. 20110005130001, the Certificate of China Postdoctoral Science Foundation Grant under grant no. 2011M500226, and the National Basic Research 973 Program of China under grant no. 2009CB320406.

7. References


