PANP-GM: A Periodic Adaptive Neighbor Workload Prediction Model Based on Grey Forecasting for Cloud Resource Provisioning

Yazhou Hu[1(✉)], Bo Deng[1], Fuyang Peng[1], Dongxia Wang[2], and Yu Yang[1]

[1] Beijing Institute of System Engineering, Beijing 100101, China
huyazhou@mail.ustc.edu.cn, {deng_bo0, fy_peng0, yu_yang0}@sina.com
[2] National Key Laboratory of Science and Technology on Information System Security, Beijing 100101, China
dx_wang0@sina.com

Abstract. Cloud computing platforms provide on-demand service to meet users' need by adding or removing cloud resources dynamically. The cloud resource provisioning is often based on the feedback model, which causes time delay and resource wasters. Workload prediction methods can make the resource provisioning more instantaneous and reduce resource and power consumption, to meet service level objectives (SLOs) and improve quality of service (QoS) of cloud platform. In this paper, we propose a periodic adaptive neighbor workload prediction model based on grey forecasting (PANP-GM) for cloud resource provisioning. Firstly, the model analyzes the growth rate and evaluates the periodicity of workload. Secondly, this model uses the growth rate of previous neighbor periodicity to predicate the trend of upcoming workload. To adapt to dynamic changes and emergencies, the grey forecasting model is applied for automatic error correction and improving prediction accuracy. Experimental results demonstrate that PANP-GM can achieve better resource prediction accuracy than basic and general approaches. Furthermore, this model can effectively improve the QoS of cloud platform and reduce SLO violations.

Keywords: Workload prediction · Periodicity · Grey forecasting model · Resource provisioning

1 Introduction

Cloud computing platform supports tenants to add or remove cloud resource dynamically based on the on-demand service. The on-demand service model provisions or removes resource depending on feedback of resource utilization. However, this method often causes over-provisioning and under-provisioning resource. Under-provisioning can't meet demands of sudden surge workload, which causes SLO violations. Over-provisioning causes resource wasters. Meanwhile, the feedback method will result in significant delays.

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 273–285, 2017.
DOI: 10.1007/978-3-319-59288-6_25

To reduce the provisioning delay and improve QoS of cloud platform, researchers propose some different workload prediction models. Based on the evaluation results, cloud platforms can provision resource to meet and match tenants' need. Although workload prediction methods have lots of advantages in resource provisioning, these methods still have many challenges such as:

(a) Complexities of workload: Cloud platform provides the variety of services and applications, which can vary different kinds and patterns of workload. Further more, the highly dynamic workload also improves the difficulty of prediction.
(b) Emergencies: Sometimes, there are certain emergencies in cloud platforms. Due to the sudden and randomness, prediction algorithms are extremely hard to predict upcoming workload based on history information.
(c) Tenants' behaviors: Different tenants have different behaviors of using cloud resource. Meanwhile, when and how to use cloud resource are preconditions of the prediction. Thus, the variety of tenants' behaviors is another challenge for workload prediction.

To counter these above prediction challenges, we propose a periodic adaptive neighbor workload prediction model based on grey forecasting for cloud resource provisioning. This model analyzes historical workload to find periodicities and characters of tenants' behaviors. Based on the periodicity of workload, this model uses the growth rate of previous neighbor periodicity to predicate the upcoming workload. Meantime, to adapt to the dynamic change and certain emergency, grey forecasting model is applied as the supplement of the prediction model. The main contributions of this paper are summarized as follows:

(a) The periodic adaptive neighbor workload prediction model based on grey forecasting is proposed in this paper. This model has high adaptability, which is suitable for periodic and dynamic workload;
(b) In addition to prediction accuracy, this paper presents QoS evaluation metrics, including lead time, under-provisioning and over-provisioning resource; Meanwhile, the relationship of time and resource is analyzed in our work;
(c) Experimental results demonstrate that PANP-GM can achieve better resource prediction accuracy than basic and general approaches and reduce SLO violations.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the problem description and background knowledge. Section 4 introduces the predicting model, PANP-GM. Section 5 describes the experiments and evaluation results. Section 6 concludes the paper along with suggestions for future research.

2 Related Work

Researchers have proposed many different prediction models for cloud resource provisioning, to meet and match tenants' need. These models can broadly divided into three categories: time series methods, machine learning methods and queuing theory methods. These methods are summarized as follows.

Time series methods are most common methods in workload prediction. Jiang et al. [1] presented an online temporal data mining system called ASAP, which was used to model and predict the cloud virtual machine demand by using moving average (MA) model et al. Khan et al. Hoffmann et al. [2] proposed a practice guide for building empirical models, such as auto regression (AR) model, to predict Apache web server workload. Morais et al. [3] proposed a framework for the implementation of auto-scaling services that followed both reactive and proactive approaches, and also proposed some predicted models, including auto correlation (AC), linear regression (LR), auto regression, auto regression integrated moving average (ARIMA) et al. [4] searched for repeatable workload patterns by exploring virtual machines, and then introduced an approach based on Hidden Markov Modeling to characterize and predict workload patterns. Roy et al. [5] developed a model-predictive algorithm for workload predicting based on auto regression moving average (ARMA). Time series approaches, such as moving average and auto regression, were also used in [6, 7].

Machine learning methods are also widely used in workload prediction. Bankole [8] developed a cloud client prediction model using three machine learning models: support vector regression, neural networks and linear regression. Imam et al. [9] presented time delay neural network and regression methods to predict the grid and cloud platform workload. Islam et al. [10] developed resource measurement and provisioning strategies using neural network and linear regression to predict upcoming resource demands.

Queuing theory methods are popular methods in workload prediction and resource provisioning. Calheiros et al. [11] presented a predicting model based on queueing network system model and QoS.

Apart from these above three main kinds of prediction methods, there some novel prediction methods are proposed. Such as, Jheng et al. [12] presented fuzzy model to prediction workload in cloud data center. Saripalli et al. [13] proposed hot spot detection for autonomic cloud computing.

Although these above methods can predict upcoming workload based on historical workload, they still have some limitations. Time series methods are easy to deploy, but the prediction accuracy is not high. Machine learning methods have better prediction accuracy, but training parameter is difficult and time-consuming. Queuing theory methods require that workload is completely random, which is difficult to achieve.

3 Preliminary

In this section, we first introduce the workload prediction problem, including prediction framework and prediction steps. Then we show cloud resource provisioning. The grey forecasting model is presented at the end of this section.

3.1 Workload Prediction

Nowadays, cloud platform, such as Amazon and Aliyun, provides elasticity mechanism which supports adding or removing cloud resource to adapt to workload changes. The elastic scaling is often based on the feedback model, which causes time delay and resource wasters. Workload prediction methods can make the resource provisioning more instantaneous and reduce resource the power consumption, to meet service level objectives (SLOs) and improve quality of service (QoS) of cloud platform. Figure 1 shows workload prediction and resource provisioning process. As shown in Fig. 1, the cloud resource provisioning mainly includes three steps, including workload analyzing, workload prediction and elastic scaling.

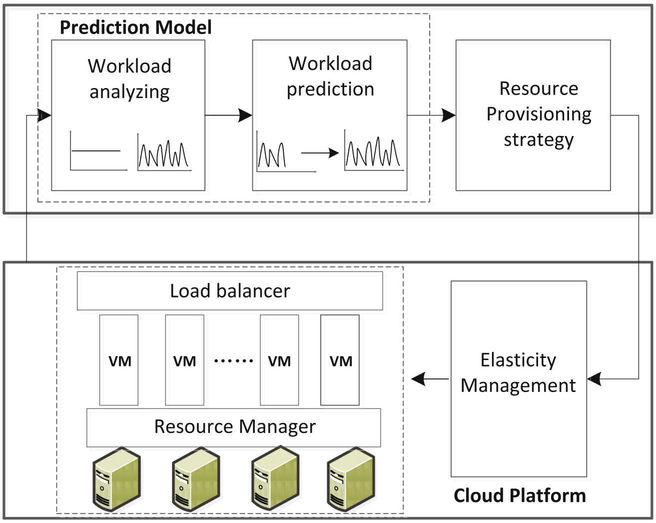

Fig. 1. Overview of cloud resource provisioning process

The prediction model is often based on historical workload data to predict trends of upcoming workload. Historical workload data include different metrics, such as CPU utilization, memory utilization, IO requests, etc. Cloud platform monitors these historical metrics and calculates tenants' resource need. Then, cloud platform scales up or scales down to provision cloud resource.

3.2 Resource Provisioning

To reduce the latency of resource allocation and meet SLOs, cloud platform should deploy resource ahead of time. Calheiros et al. [11] introduced that cloud provisioning consists of three key steps: virtual machine provisioning, resource provisioning and application provisioning. Virtual machine provisioning involves instantiation of virtual machines to coordinate hardware and software. Resource provisioning is the managing and scheduling of virtual machines on physical servers. Application provisioning is deployment of predictions to meet the SLOs. In commercial cloud platform, cloud resource provisioning is often based on virtual machines.

3.3 Grey System Theory

Grey forecasting model is based on the grey system theory [14] and fuzzy model. Deng J. L. proposed the grey system in 1982. This system can deal with the fuzzy and incomplete data. To update and complete the uncertain data, grey system uses related analysis methods, such as modeling, forecasting, and etc.

The grey forecasting model is the most common method in grey system theory. It can predict uncertain factors and tendency of case based on the history data. The grey forecasting model is often based on the differential equation, and $G(1,1)$ is the fundamental model of grey forecasting model.

4 Prediction Model

In this section, we first introduce the overview of our prediction model, including assumptions and prediction steps. Then we introduce these prediction algorithms and models in detail, which includes workload analysis algorithm, grey forecasting model and prediction algorithm.

4.1 Model Overview

The prediction model, PANP-GM, collects the historical workload, and then analyzes the characteristics of workload, to predict trends of upcoming workload. Meanwhile, two foundational assumptions are made by PANP-GM. The first one is that the workload is periodic fluctuation. The second one is that workload can't be too heavy so that cloud servers go down. In other words, the value of workload conforms to normal distribution law. In our research, we find that most of workload is in accordance with above two foundational assumptions, thus our prediction model is universal and efficient for most of workloads.

PANP-GM has two main step, workload analysis and workload prediction. For the first step, the proposed model analyzes the collected workload data, calculates growth rate of workload and estimates the period value of workload. The first step is based on the workload analysis algorithm. For the second step, the proposed model uses the growth rate of previous neighbor workload periodicity, which is proposed in first step, to predicate the upcoming workload. Meantime, to adapt to dynamic changes and abnormal increase and decrease, the grey forecasting model is applied for automatic error correction and improving prediction accuracy. To simplify the complexity of calculation, our prediction model use $G(1,1)$ as the grey forecasting model.

4.2 Prediction Algorithms

As described in above subsection, the prediction model includes three main algorithms and models, workload analysis algorithm, grey forecasting model and prediction algorithm.

(1) Workload analysis algorithm

The workload analysis algorithm is shown as Algorithm 1. Firstly, it calculates growth rate of workload data, V_o. In our research, the metabolic range of original workload data is high. However, the growth rate of workload data can reduce the impact of amplitude variation. So, this algorithm estimates the period value of workload based on analyzing V_o in the next step. In this algorithm, we use a high-efficiency period detection algorithm, ERPP, which is based on edit distance with real penalty and proposed in reference [15].

Algorithm 1. Workload analysis algorithm

Inputs:

X_o : original workload data

Outputs:

V_o : the growth rate of workload data

t : the period value of workload

```
Intialize V_o = 0, t = 1
for i=2 to n
    V_o(i) = (X_o(i) - X_o(i-1)) / X_o(i-1)
end for
call algorithm ERPP to calculate the candidate value p of V_o
t ← p
return t
```

(2) Grey forecasting model

Grey forecasting model predicts the workload tendency based on the exiting workload data. $G(1,1)$ is the basic one of grey forecasting model, which includes three main steps described as follows.

(i) accumulation generating operation

For the original workload data $X^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots x^{(0)}(n))$, the model uses one time accumulation generating operation to smooth the original data. The operation is shown as formula 1.

$$\begin{cases} X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots x^{(1)}(n)) \\ x^{(1)}(k) = \sum_{1}^{k} x^{(0)}(k), k = 1, 2, \cdots, n \end{cases} \tag{1}$$

Algorithm 2. Prediction algorithm

Inputs:

$X^{(0)}(k), k = 1, 2, \cdots m$: original workload sequence

α : the threshold of calling grey forecasting model

β : the original workload sequence length of $G(1,1)$

Outputs:

$\hat{X}^{(0)}(k), k = m+1,, \cdots n$: the predictive workload sequence

```
Intialize α , β
call Algorithm 1 to calculate the period value of workload and V_o
for i=m+1 to n
    φ_o(i−1) = (x̂^(0)(i−1) − x^(0)(i−1)) / x^(0)(i−1)
    V_o(i−1) = (x^(0)(i−1) − x^(0)(i−2)) / x^(0)(i−2)
    if φ_o(i−1) > α
        call G(1,1) to predict x̂^(0)(i) based on (x^(0)(i−1−β),···,x^(0)(i−1))
    else
        x̂^(0)(i) = x^(0)(i)·(1+V_o(i−t))
    end if
end for
return X̂^(0)(k)
```

(ii) modeling

After getting data array $X^{(0)}$ and $X^{(1)}$, this model constructs a mean generating sequence $Z^{(1)}$, which is described as formula 2. Then, the $G(1,1)$ model can be described as formula 3, where a and b are adjusting factors.

$$\begin{cases} Z^{(1)} = (z^{(1)}(2), z^{(1)}(3), \cdots x^{(1)}(n)) \\ z^{(1)}(k) = 0.5\,(x^{(1)}(k) + x^{(1)}(k-1)), k = 2, 3, \cdots, n \end{cases} \tag{2}$$

$$x^{(0)}(k) + az^{(1)}(k) = b, k = 2, 3, \cdots, n \tag{3}$$

(iii) solution

To get the predicting results, this model use least-square estimation method to solve the formula 3. The predictive result $\hat{x}^{(1)}(k)$ is presented in formula 4.

$$\hat{x}^{(1)}(k) = (x^{(1)}(1) - \frac{b}{a})e^{-at} + \frac{b}{a}, k = 1, 2, \cdots, n \quad (4)$$

So, the predictive value is described as formula 5.

$$\hat{x}^{(0)}(k+1) = (1 - e^{a})(x^{(0)}(1) - \frac{b}{a})e^{-ak}, k = 1, 2, \cdots, n \quad (5)$$

Thus, the predictive sequence is $\hat{X}^{(0)} = (\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \cdots \hat{x}^{(0)}(n))$.

(3) **Prediction algorithm**

The prediction algorithm is the key factor of grey forecasting model. This algorithm uses the growth rate of previous neighbor periodicity as the present growth rate to predicate the upcoming workload. If the predicted data has a big difference with the real data, the prediction algorithm applies $G(1,1)$ for automatic error correction and improving prediction accuracy. The difference percentage is described as $\varphi_o(k)$. The coefficient α is the threshold of calling grey forecasting model and β is the train sequence length of $G(1,1)$. This algorithm is shown as Algorithm 2.

5 Performance Evaluation

This section presents results evaluating PANP-GM model. We first introduce the experiment setup and evaluation metrics. Then, we tune parameters of prediction model. At last, we present and analyze experiment results.

5.1 Experiment Setup

In our experiment, we use the traffic of data center as the workload, and we monitor and collect workload data for 456 h. The PANP-GM model is applied in workload prediction based on the collected data.

To evaluate the performance of our model, we compare our model with three baseline methods, including AR, MA and ARIMA. Meantime, we propose two different kinds of evaluation metrics, which includes prediction error and QoS evaluation metrics. These metrics are shown as follows.

(a) **Prediction error**

The mean of absolute percentage error (MAPE) is used to evaluate the prediction accuracy of PANP-GM model. The metric is defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{x}^{(0)}(i) - x^{(0)}(i)}{x^{(0)}(i)}\right| \quad (6)$$

Where, $x^{(0)}(i)$ is the original workload value and $\hat{x}^{(0)}(i)$ is the predicted workload value.

(b) **Qos evaluation metrics**

Time and resource are two important factors affecting the quality of service. To evaluate the improvement of QoS and reducing of SLO violations, we propose three QoS evaluation metrics, including lead time rate, under-provisioning and over-provisioning resource rate.

Leading time rate: leading time is the proportion of saved time obtained by the prediction model, which is defined as formula 8.

$$R_t = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{t_i - t_{pi}}{t_i} \tag{7}$$

t_i is the provisioning time without prediction, and t_{pi} is the provisioning time using the prediction framework.

Under-provisioning and over-provisioning resource rate: Under-provisioning can't meet demands of sudden surge workload, which causes SLO violations. Over-provisioning causes resource wasters. Thus, these metrics are considered for evaluate the prediction model. The under-provisioning resource (R_u) and over-provisioning resource (R_o) rates are defined as follow:

$$R_u = \frac{1}{n}\sum\nolimits_{t=1}^{n} \frac{R_d - R_p}{R_d}, R_d \geq R_p \tag{8}$$

$$R_o = \frac{1}{n}\sum\nolimits_{t=1}^{n} \frac{R_p - R_d}{R_d}, R_p > R_d \tag{9}$$

R_d is the demand resource of workload, and R_p is the prediction resource.

5.2 Parameters Turning

There are some parameters in our prediction algorithms should be adjusted to improve the prediction performance. So, we train and select optimal parameters based on experimental analysis.

(a) **Threshold and sequence length**

There are two parameters, α and β, in Algorithm 2. To improve the performance of prediction method, these parameters should be turned for getting the exact value. In this experiment, we change the values of parameters to select the optimal parameters.

α is the threshold of calling grey forecasting model, and β is the train sequence length of $G(1,1)$. To get optimal values of α and β, we analyze the relations between and among threshold, sequence length and prediction error. As shown in Fig. 2, the prediction error increases first, and then decreases with the increase of threshold value.

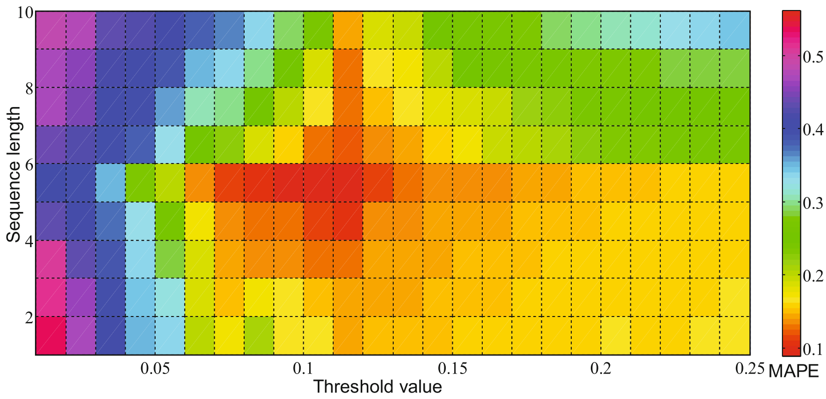

Fig. 2. Relations between and among threshold, sequence length and prediction error

Meantime, we can find that the prediction error is optimal when values of α and β are 0.11 and 5. Thus, we select these values to predict upcoming workload in the next experiments.

(b) **Period value**

To predict the trends of upcoming workload, we should analyze the periodicity of workload firstly. As shown in Fig. 3, the growth rate is more obvious than original workload in periodic changes. Therefor, we use the growth rate series to predict the trends of upcoming workload. Figure 3 shows that the period value is 24.

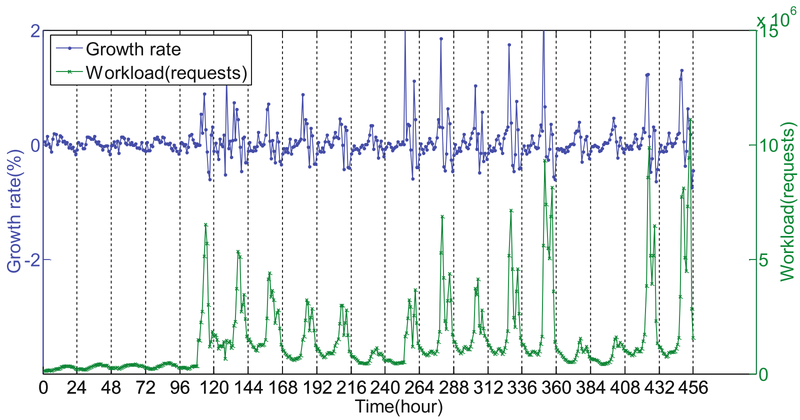

Fig. 3. Workload and growth rate analyzing

(c) **Prediction steps**

Algorithm 2 supports multi-step prediction. At the same time, prediction steps and leading time are equivalent in our algorithm, thus we should analyze the relations between and among leading time, prediction error, under-provisioning resource rate and over-provisioning resource rate. As shown in Fig. 4, with the leading time (prediction steps) increasing, the prediction error, under-provisioning resource rate and over-provisioning resource rate are increasing rapidly. So, synthesizing these results, we select one-step prediction in Algorithm 2.

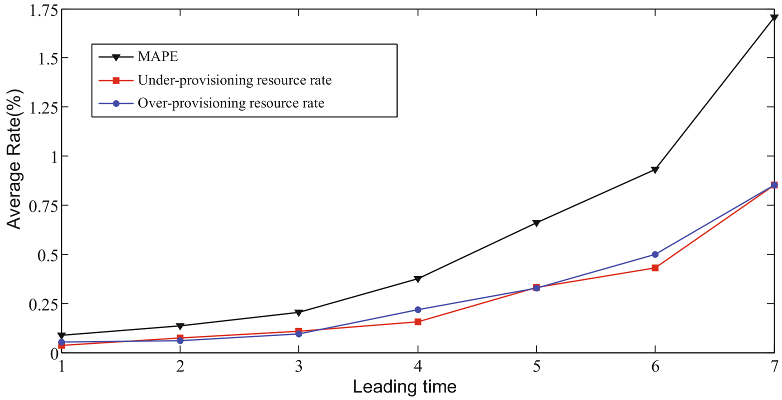

Fig. 4. Prediction performance varies along leading time

5.3 Experiment Results

Based on the above subsection analysis, we choose optimal parameters and conduct experiments using the prediction algorithm. Moreover, we also compare our model with three basic methods, including AR, MA and ARIMA. Experiment results are shown in Table 1 and Fig. 5.

As shown in Fig. 5 and Table 1, MAPE, under-provisioning resource rate and over-provisioning resource rate of our prediction method, PANP-GM, are significantly less than other three basic methods. The prediction error is less than 10%, and the under-provisioning resource rate and over-provisioning resource rate are above 5%. Further more, the leading time rate is more than 80%, which is more efficient than other methods.

Based on these experiment results, our method can achieve better prediction accuracy and significantly reduce SLO violations than other three basic methods. In other words, PANP-GM model can improve QoS of Cloud platform by workload prediction for resource provisioning.

Table 1. Evaluation results of different prediction methods

Evaluation metrics/methods	AR	MA	ARIMA	PANP-GM
MAPE	0.268	0.287	0.223	0.089
R_t	0.531	0.572	0.636	0.891
R_u	0.106	0.149	0.103	0.037
R_o	0.162	0.138	0.120	0.052

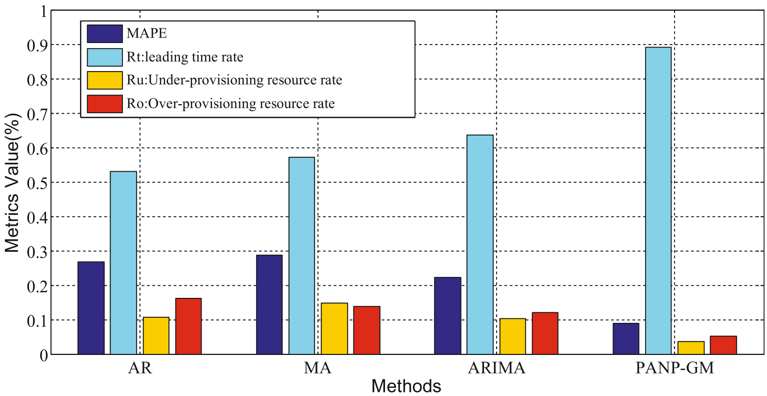

Fig. 5. Comparison of prediction performance of different methods

6 Conclusion and Future Work

This paper proposes a periodic adaptive neighbor workload prediction model based on grey forecasting (PANP-GM) for cloud resource provisioning. This model uses the growth rate of previous neighbor periodicity to predicate the trend of upcoming workload based on workload analysis. Furthermore, the grey forecasting model is applied for automatic error correction and adapting dynamic changes and emergencies. Experimental results demonstrate that PANP-GM can achieve better resource prediction accuracy than AR, MA and ARIMA model. Meanwhile, this model can effectively improve the QoS of cloud platform and reduce SLO violations.

In this paper, the model proposed achieves poor performance on irregular workload prediction and multi-step prediction. Therefore, efficiently dynamic and adaptive prediction methods for complex workload should be proposed. At the same time, the elastic scaling strategy based on the prediction results should be focused on in the future work.

Acknowledgments. This work is supported by the National High-Technology Research and Development Program of China (863 Program) (No. 2013AA01A215) and the National Natural Science Foundation of China (No. 61271252).

References

1. Jiang, Y., Perng, C., Li, T., et al.: Asap: a self-adaptive prediction system for instant cloud resource demand provisioning. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 1104–1109. IEEE (2011)
2. Hoffmann, G., Trivedi, K.S., Malek, M.: A best practice guide to resource predicting for computing systems. IEEE Trans. Reliab. **56**(4), 615–628 (2007)
3. Almeida Morais, F.J., Vilar Brasileiro, F., Vigolvino Lopes, R., et al.: Autoflex: service agnostic auto-scaling framework for iaas deployment models. In: 2013 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 42–49. IEEE (2013)
4. Khan, A., Yan, X., Tao, S., et al.: Workload characterization and prediction in the cloud: a multiple time series approach. In: 2012 IEEE Network Operations and Management Symposium (NOMS), pp. 1287–1294. IEEE (2012)
5. Roy, N., Dubey, A., Gokhale, A.: Efficient autoscaling in the cloud using predictive models for workload predicting. In: 2011 IEEE International Conference on Cloud Computing (CLOUD), pp. 500–507. IEEE (2011)
6. Reig, G., Guitart, J.: On the anticipation of resource demands to fulfill the QoS of saas web applications. In: Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing, pp. 147–154. IEEE Computer Society (2012)
7. Xu, W., Zhu, X., Singhal, S., et al.: Predictive control for dynamic resource allocation in enterprise data centers. In: 10th IEEE/IFIP Network Operations and Management Symposium, NOMS 2006, pp. 115–126. IEEE (2006)
8. Bankole, A.A., Ajila, S.A.: Cloud client prediction models for cloud resource provisioning in a multitier web application environment. In: 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), pp. 156–161. IEEE (2013)
9. Imam, M.T., Miskhat, S.F., Rahman, R.M., et al.: Neural network and regression based processor load prediction for efficient scaling of Grid and Cloud resources. In: 2011 14th International Conference on Computer and Information Technology (ICCIT), pp. 333–338. IEEE (2011)
10. Islam, S., Keung, J., Lee, K., et al.: Empirical prediction models for adaptive resource provisioning in the cloud. Future Gener. Comput. Syst. **28**(1), 155–162 (2012)
11. Calheiros, R.N., Ranjan, R., Buyya, R.: Virtual machine provisioning based on analytical performance and QoS in cloud computing environments. In: 2011 International Conference on Parallel Processing (ICPP), pp. 295–304. IEEE (2011)
12. Jheng, J.J., Tseng, F.H., Chao, H.C., et al.: A novel VM workload prediction using Grey Predicting model in cloud data center. In: 2014 International Conference on Information Networking (ICOIN), pp. 40–45. IEEE (2014)
13. Saripalli, P., Kiran, G.V.R., Shankar, R.R., et al.: Load prediction and hot spot detection models for autonomic cloud computing. In: 2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC), pp. 397–402. IEEE (2011)
14. Julong, D.: Introduction to grey system theory. J. Grey Syst. **1**(1), 1–24 (1989)
15. Wang, Y., Gao, X., Wu, S., et al.: Periodicity detection method of periodic pattern mining in time series. Comput. Eng. **22**, 014 (2009)