Research on the Prediction Model of CPU Utilization Based on ARIMA-BP Neural Network

Jina Wang[1], Yongming Yan[2,a] and Jun Guo[2]

[1] *Liaoning Software Testing Center, Shenyang, China*
[2] *College of Information Science & Technology Engineering, Northeastern University, Shenyang, China*

Abstract. The dynamic deployment technology of the virtual machine is one of the current cloud computing research focuses. The traditional methods mainly work after the degradation of the service performance that usually lag. To solve the problem a new prediction model based on the CPU utilization is constructed in this paper. A reference offered by the new prediction model of the CPU utilization is provided to the VM dynamic deployment process which will speed to finish the deployment process before the degradation of the service performance. By this method it not only ensure the quality of services but also improve the server performance and resource utilization. The new prediction method of the CPU utilization based on the ARIMA-BP neural network mainly include four parts: preprocess the collected data, build the predictive model of ARIMA-BP neural network, modify the nonlinear residuals of the time series by the BP prediction algorithm and obtain the prediction results by analyzing the above data comprehensively.

1 Introduction

In the recent cloud computing research field, the virtual resource management problem has become a hotspot. The VM dynamic deployment technique is one of the key virtual resource management technique. Many experts have done a lot of research about it.

At present, the study of the VM resource dynamic deployment focus on the fine-grained resource adjustment strategy. The current workload data and the future change trend data are needed if a predictive and accurate resource adjustment plan is to be made. Daniel A Menasce[1] put forward a dynamic resource adjustment plan following the changing workload by the CPU priority or the CPU shares. ZHAO Weiming[2] presented a new theory to improve memory utilization by predicting memory usage of each VMs. Anton Beloglazov[3] proposed a set of the heuristic algorithms to select the objective VM for migration when the VM violate the SLA. A new distributed management strategy select the VM whose CPU usage rate is the maximum in the abnormal server in the reference 4. If after migrating the CPU usage rate still be high, it would continue to migrate the VM whose CPU usage rate is the maximum in the rest VMs until recovering the server. Compared with the reference 4, some researchers put forward a new algorithm to select the VM for migration by predicting the workload trend so that it can avoid the instantaneous peak load [5].

2 The Research Method of the CPU Utilization Prediction Process

With the ongoing changes of the workload, the service performance should be ensured by adjusting the resource of the VM. After adding the predictive method, the VM dynamic deployment plan can avoid the performance fluctuation caused by the lag of the optimization process. By the historical data a predictive model of the future CPU utilization can be set up. With the predictive information, more supports can be received by the dynamic deployment plan.

So far, the method of the weighted average to predicted single resource change trend is widely used in the existing models, but the method using the composite model to predict the CPU utilization by the time series analysis technique is adopted in few researches. The research emphasis in the former is to design the weighting efficient system. A good weighting efficient system means an accurate predicted result. But with much subjectivity and arbitrariness it is very difficult to design a good system. By researching, it is found that a time series process consists of a linear structure and a nonlinear structure that means the change trend rule can be found easily by the time series.

The linear prediction model is mainly adopted in the time series prediction method so that the method cannot process the nonlinear data well. Compared with the traditional data processing algorithms, BP neural network is an effective nonlinear modeling method. It has great

[a] Corresponding author: yym_sy@163.com

advantages in processing the feature-undefined data that with much randomness, nonlinearity and noisy. Implied non-linear relationships can be excavated out from the database by the BP model. But it is weaker in process the linear data than the traditional methods such as time series model. In fact most server CPU utilization sequences usually consist of linear time sequence part and nonlinear time sequence part that means it mix the character of linear and nonlinear. So the composite model combining with BP neural network and time series prediction method can improve the predictive result.

According to the analysis, a new predictive model combing with ARIMA model and BP neural network to predict the future CPU utilization is put forward in this paper.

The CPU utilization is fitted by the ARIMA prediction technique in the new model, then the nonlinear residual is calculated by the BP neural network and finally the real predictive result is obtained by adding the both parts. The relevant technical theories are introduced as follows.

2.1 BP Neural Network

BP(Back propagation), an abbreviation for "backward propagation of errors", is a common method of training artificial neural networks used in conjunction with an optimization method such as gradient descent.

Each propagation involves the following steps:

(1) Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.

(2) Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas of all output and hidden neurons.

For each weight-synapse follow the following steps:

(1) Multiply its output delta and input activation to get the gradient of the weight.

(2) Subtract a ratio (percentage) of the gradient from the weight.

2.2 Time Series Prediction

The steps of building the ARIMA(p,d,q) model are shown as follows:

(1) Collect the time series data set;

(2) Draw out the sequence diagrams, autocorrelation function diagram and partial autocorrelation function diagrams of the observed data. And judge the series whether is stationary series or not. If not it should be process by the differential computing process to get the stationary series.

(3) Calculate the Q_{LB} statistics of the observed series, and verify the stochastic feature of the series.

(4) The ARIMA(p,d,q) model is built by combine the feature values of the AR function and the partial AR function. Select a proper order value and then estimate model parameters.

(5) Calculate the Q_{LB} statistics of the residual sequence. Only when the residual sequence is the white noise sequence the built-model is a effective model.

(6) By the forecasting model the prediction values that is the important data for us to make decisions can be obtained.

3 The Modeling Method of the ARIMA-BP Model

3.1 Building the ARIMA(p, d, q) Model.

Step1：Collect the server CPU utilization data. And mark them as original series after preprocess them.

$$y\left(t\right) = \left\{y_1, y_2, \ldots, y_n\right\} \qquad (1)$$

Step2：The original series should be smoothened. In this paper the autocorrelation is the index to measure the degree of tranquilization.

The autocorrelation is the ratio of the covariance of the h-lag time series to the total variance of the time series: $\left(\text{ACs}\right) \rho_h = \gamma_h / \sigma^2$. Where the γ_h can be obtain by follow formula.

$$\gamma_h = \frac{1}{T}\sum_{t=h+1}^{T}\left(y_t - \bar{y}\right)\left(y_{t-h} - \bar{y}\right) \qquad (2)$$

Where the $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$ is the sample meas. The autocorrelation descript the relation between the time series values and its h-lag period. One of the most important characteristic for the stable sequence is that the autocorrelation function ρ_h is rapidly descend to 0 by the increase of the h-lag period. For the unstable time series the union-change will be week.

If the time series can be proved to be unstable, its unstable features can be separate out so that the stable time series can be obtained. In this paper the method that combining the logarithm transformation and the difference algebra is used to smoothen the time series.

Step3：Calculate the autocorrelation coefficients and the part autocorrelation coefficients. And then build the time series model by the identification rule. After calculating, the autocorrelation coefficients and the part autocorrelation coefficients should follow the ARMA(p,q). Finally the order parameters will be make sure by the AIC rules. In this paper we calculate the order values are p=1，q=1. Because the time series is just with the first difference, the final model form is ARIMA（1，1，1）

Step4 : It is very significant to estimate the parameters. After getting the stable time series, the follow method will be used to build the ARMA(p, q) model:

$$u_t = y_t - \varphi y_{t-1} - \cdots - \varphi_p y_{t-p} + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \qquad (3)$$

Work out the $u_t = u_t\left(\varphi, \theta\right)$. Where the t=p+1,…,T. The purpose of the parameter estimation is to get the min sum square by finding a set of parameters $\left(\varphi_1, \varphi_2, \ldots, \varphi_p, \theta_1, \theta_2, \ldots, \theta_q\right)$

$$S\left(\varphi, \theta\right) = \sum_{t=p+1}^{n} u_t^2\left(\varphi, \theta\right) \quad (4)$$

Step5： And then the model need to be checked. Calculate the residual sequence. The model can be proved to be accuracy only when the residual sequence is white sequence.

Step6：Predicting the CPU future utilization by the change rules finding by the fitting model. The predicted data are the key reference for the future decisions.

3.2 Modifying the Residual Error Correction by the BP Neural Network

Step1：Calculate the residual sequence of the CPU utilization time series.

$$N_i = y_i - L'_{i'}, \left(i = 1, 2, 3, \ldots, n\right) \quad (5)$$

Where the $\left\{L'_1, L'_2, \ldots, L'_n\right\}$ is the predicted server CPU utilization data getting by the ARIMA(p, d, q) model. Where the $\{y_1, y_2, \ldots, y_n\}$ is the real server CPU utilization data.

Step2：Normalize the residual sequence so that is can be process in the sigmoid function.

$$N_i = \left(N_i - s\right)/\left(t - s\right) \quad (6)$$

Where the $s=(9(N_i)_{min}-(Ni)_{max})/8$ and $t=(9(N_i)_{max}-(Ni)_{min})/8$. Where the $(Ni)_{max}$ is the maximum value of the residual sequence and where the $(Ni)_{min}$ is the minimum value of the residual sequence.

Step3：Define the structure of the BP neural network

Analyzing the residual sequence of the server CPU utilization is the main function of the neural network. In this paper the input structure is single-input that means the number of the input layer neuron is 1 and the same with the number of the output layer neuron. The number of the hidden neurons is define as follow formula:

$$l = \sqrt{\left(m + n\right)} + a \quad (7)$$

Where the m is the number of the input layer. Where the n is the number of the output layer. Where the a is a random constant range from 0 to 10.

Step4： Define the trained parameters of the BP neural network： The main parameters of BP neural network are the transfer function of the hidden layer, the transfer function of the output layer, the learning rate and the iterations.

Step5：The neural network need to be trained several times. The different network need to be compared so that the best one can be selected out. And at last the residual sequence N'_t can be obtained by the network simulation.

3.3 The prediction of the server CPU utilization

The final predicted data of the server CPU utilization y'_t that based on the ARIMA-BP neural network algorithm can be obtained by adding the initial predicted data of the server CPU utilization y'_t predicted by the ARIMA(p,d,q) and the residual sequence N'_t predicted by the BP neural network.

4 Analysis of Experimental Results

4.1 The experimental steps

(1) Filter the Data.

After be preprocessed the data of the CPU utilization monitored by the monitoring program is the original time series of the server CPU utilization. The interval is 15 seconds. Select 300 records with the interval to predict the CPU utilization in future 30 intervals. The time series of the server CPU utilization is shown in Figure.1.

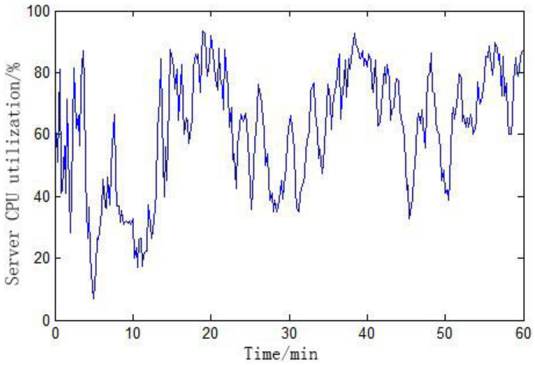

Figure 1. CPU utilization of physical server.

(2) Smoothen the original time series.

In this paper the method that combining the logarithm transformation and the difference algebra is used to smoothen the time series. The processed result is shown in Figure. 2.

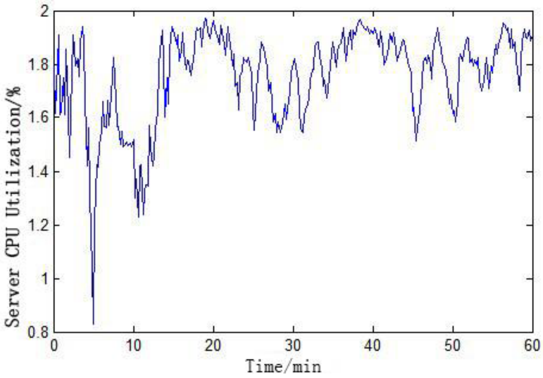

Figure 2. After taking logarithm of the CPU utilization of physical server.

(3) Built the ARIMA(p, d, q) model.

The method of building the model is combined the contents in the chapter 4.4.2 and the AIC rules. It is obviously when R=1 and M=1, the values of AIC and BIC can be got the minimum. So the model ARIMA(1,1,1) is chosen in the experiment. After be checking the residual sequence of the model can be proved to be white noise sequence that means the model is effective.

R=0,M=0,AIC=2202.480317,BIC=2209.881204
R=0,M=1,AIC=2202.148436,BIC=2213.249767

R=0,M=2,AIC=2200.471689,BIC=2215.273463
R=0,M=3,AIC=2200.087553,BIC=2218.589771
R=1,M=0,AIC=2202.601972,BIC=2213.703303

(4) Modify the residual sequence by the BP neural network

The residual sequence of the predicted value need to be worked out t and analyzed. And then the residual sequence need to be modified. The final predicted data of the server CPU utilization can be obtained by adding the initial predicted data of the server CPU utilization predicted by the ARIMA(1,1,1) and the analyzed result of the residual sequence.

4.2 The analysis of experiment results

The details of the IBM server is shown as follows: eight-core(two Intel Xeon E5506 2.13GHz), 16 GB 2-channel memory (1333MHz), 500G hard disk and with Linux system. For to creating and managing five virtual machines, the Xen system is installed in the server. And the Xen system is also used to manage the resource of the server. The resource of the Xen's virtual domains should be allocated when starting the virtual domains. The physical resource of the server will be shared by all the virtual domains in the form of partitions. All the utilization of the server resource will be monitored by the Xen system.

The composite method ARIMA(p,d,q)-BP is compared to the signal method ARIMA(p,d,q) and the signal method BP neural network. From the following data figure and tables it can be seen that the result that come from the composite model is more close to the real value.

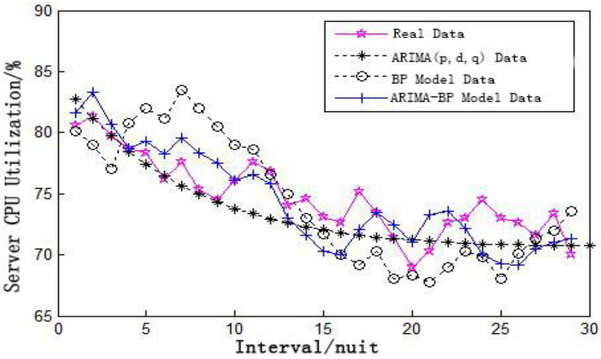

Figure 3. The Comparative Result between Model Value and Real Value of Three Predicted.

Table 1. The Errors Comparative Result of Three Predicted Algorithm.

Algorithm	mean relative error	mean square error	mean absolute error
ARIMA	0.0370	0.303	0.208
BP	0.0265	0.228	0.125
ARIMA -BP	0.0186	0.126	0.097

After Analyzing the data in table 1, it can be seen that the ARIMA(p,d,q)-BP model is more accurate than the single model. The mean relative error separately decreased from 0.0370 and 0.0265 down to 0.0186. The mean square error separately decreased from 0.303 and 0.0228 down to 0.0126. The mean absolute error separately decreased from 0.208 and 0.125 down to 0.097. So the composite model is proved to be more accurate.

5 Conclusion

In this paper, the new composite model combining with the ARIMA(p,d,q) model and the BP neural network is proposed. With the composite model, the future change trend rules of the server CPU utilization can be found. The prediction value of the future server CPU utilization predicted by the new model can be more close to the real value. With the accurate predictive data the strategy of the VM resource dynamic deployment will be more useful.

Acknowledgments

This work was supported by the National Natural Science Foundation Program of China (61572117, 61300019, 61370155) and the Special Fund for Fundamental Research of Central Universities of Northeastern University (N120804001, N120204003, N140406002).

References

1. Menasce D A, Bennani M N. Autonomic virtualized environments[C].Autonomic and Autonomous Systems, 2006. ICAS'06. 2006 International Conference on. IEEE, 2006: 28-28.

2. Zhao W, Wang Z, Luo Y. Dynamic memory balancing for virtual machines[J]. ACM SIGOPS Operating Systems Review, 2009, 43(3): 37-47.

3. Beloglazov A, Abawajy J, Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing[J]. Future Generation Computer Systems 2011; doi:10.1016/j.future.2011.04.017.

4. MA Fei, LIU Feng, LIU Zhen Efficiency Energy-saving Distributed Management Method of Virtual Machine in Cloud Computing [J]. Computer Engineering, 2012, 38(11).

5. HU Zhi-gang, OUYANG Cheng, YAN Chao-kun Resource Load Balancing Method for Energy-consumption Reducing in Cloud Environment[J] Computer Engineering, 38(5): 53-55.