Load prediction of grid computing resources based on ARSVR method

Huang Gang Wang Ruchuan Xie Yongjuan Shi Xiaojuan

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Based on the monitoring and discovery service 4 (MDS4) model, a monitoring model for a data grid which supports reliable storage and intrusion tolerance is designed. The load characteristics and indicators of computing resources in the monitoring model are analyzed. Then, a time-series autoregressive prediction model is devised. And an autoregressive support vector regression (ARSVR) monitoring method is put forward to predict the node load of the data grid. Finally, a model for historical observations sequences is set up using the autoregressive (AR) model and the model order is determined. The support vector regression (SVR) model is trained using historical data and the regression function is obtained. Simulation results show that the ARSVR method can effectively predict the node load.

Key words: grid; autoregressive support vector regression algorithm; computing resource; load prediction

In order to achieve effective organization of grid resources, provide convenient user access and maximize resource utilization, the effective maintenance and management of a system and its resources becomes an important aspect for a grid system, which requires an adaptive monitoring system. The monitoring system not only protects the normal operation of the grid system, but also provides state information resources for other services. An appropriate grid monitoring system is urgently needed to accurately access real-time grid information, which can provide a data foundation for resource scheduling and performance optimization of the data grid.

The traditional grid resource monitoring systems are mostly developed for the existing grid systems and difficult to be transplanted to other systems. The grid resource monitoring systems can only monitor the load of real-time nodes but do not make effective predictions. This monitoring pattern cannot meet the needs of grid resource scheduling and performance optimization.

In this paper an autoregressive support vector regression (ARSVR) method is proposed for a data grid which supports reliable storage and data intrusion. Simulation results show that the method can predict the node load in real time and provide a basis for resource scheduling and other grid services.

1 Monitoring Model

The open grid service architecture (OGSA) is a service-oriented grid architecture, which is built on the basis of grid services. In the OGSA, all are abstracted as services, including computing resources, storage resources, networks, programs, databases, instruments and equipment, etc[1]. The OGSA takes the grid service as the core and provides all kinds of services for users through the interface of the grid service. The grid service is composed of service data and achievement. On the basis of the OGSA, a data grid which supports reliable storage and intrusion tolerance is constructed.

Monitoring and discovery service 4 (MDS4), a widely-used monitoring system, provides a framework which can manage and compute the dynamic and static information of a grid. The functions of MDS4 include discovering resources and providing state information, resource scheduling and monitoring information[2]. Monitoring real-time information of resources is not enough. For a data grid which supports reliable storage and intrusion tolerance, a suitable monitoring system is needed to provide a global and abstract resource view. The system should enable users to quickly monitor data file information. Under the requirement of real-time monitoring of computing resources, a prediction-based method for a performance monitoring system should be proposed to predict the node load of a data grid, which can choose appropriate grid nodes for different tasks and provide data for resource scheduling and performance optimization. On the basis of the MDS4 model, a monitoring model based on a data grid which supports reliable data storage and intrusion tolerance is proposed, as shown in Fig. 1.

The monitoring model based on a data grid is divided into three levels: resource layer, service layer and application layer. The resource layer is responsible for taking a variety of heterogeneous resources to access grid systems. Resource information is gathered and resource interfaces are packaged in the layer. The service layer works under the GT4 container environment. It makes use of various information providers to collect different kinds of grid resources through various sources, and congregates various services to provide a unified query interface for the client. The application layer provides application program interfaces, command lines and graphical interfaces to interact with the intermediate service layer.

2 Load Prediction

In the data grid which supports reliable storage and intrusion tolerance, due to the autonomy of nodes, resources are not subject to the control of the grid, which makes task scheduling and performance optimization difficult. Therefore, the data grid environment not only monitors the status of current resources, but also grasps the load characteristics of grid nodes, which can provide a reliable reference for

Received 2009-04-30.

Biography: Huang Gang(1961—), male, associate professor, huanggang@njupt.edu.cn.

Foundation item: The National High Technology Research and Development Program of China (863 Program) (No. 2007AA01Z404).

Citation: Huang Gang, Wang Ruchuan, Xie Yongjuan, et al. Load prediction of grid computing resources based on ARSVR method[J]. Journal of Southeast University (English Edition), 2009, 25(4): 451 –455.

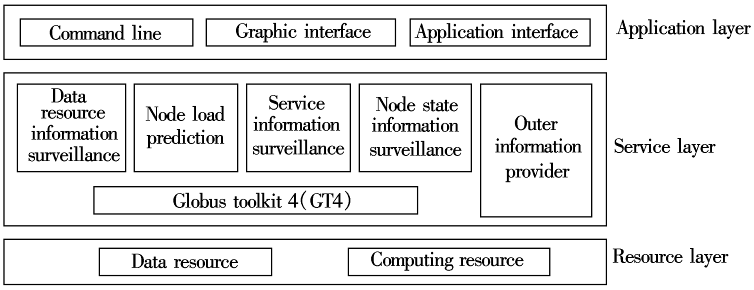

Fig. 1 Monitoring model based on a data grid which supports reliable storage and intrusion tolerance

task scheduling. By accurately measuring the host load and finding the discipline of load changes, a feasible degree of host load prediction can be obtained[3]. The analysis results of many load patterns show that host load changes are complicated and some disciplines can be obtained. It is entirely feasible to accurately predict the load by some appropriate methods.

In the process of load prediction, the selection of load indicators is important. Factors directly related to a machine's load are called load indicators, among which some can be used to measure the load of the host. The utilization rate of system resources can be used as a parameter for evaluating the load. This method ignores the type of the task. For system performance evaluation, the resources of the host can be divided to the following parts: CPU utilization rate, I/O utilization rate, bandwidth utilization rate and memory utilization rate. When evaluating the performance of a system, a weight value for every resource should be set[4]. The weight value can be determined by the tasks running in the system. The integrated load L of a grid node can be calculated by

$$L = \alpha L_{\mathrm{C}} + \beta L_{\mathrm{I}} + \gamma L_{\mathrm{B}} + \mu L_{\mathrm{M}} \tag{1}$$

where α, β, γ and μ are the weight values of the CPU utilization rate, the I/O utilization rate, the bandwidth utilization rate and the memory utilization rate, respectively.

However, α, β, γ and μ are difficult to be determined. We can measure n groups of L_{C}, L_{I}, L_{B} and L_{M} at different times. Then, the weight value of each parameter can be calculated by

$$W_i = \frac{L_i}{\sum_{i=1}^{n} L_i}$$

where W_i is the weight of the i-th parameter; L_i is the utilization rate of the i-th parameter.

The average values of L_{C}, L_{I}, L_{B} and L_{M} can be separately obtained. If the testing data are enough, the results can be used as the average of the CPU utilization rate, the I/O utilization rate, the bandwidth utilization rate and the memory utilization rate in normal scenarios. Then, the average values are summed. The ratio of each average to the sum is taken as a weight value for predicting performance evaluation. Under normal circumstances, a high utilization rate has a high weight, which is conducive to reflect the system utilization rate. In most cases, the results are stable.

3 ARSVR Load Prediction Method

3.1 Prediction analysis

According to the characteristics of a host load, load changes can be regarded as a time-series process, which has a strong correlation with time changes. It is appropriate to use a time-sequence method for host load prediction. The time sequence contains the observed sequence values of a variable y at different times t and can be expressed as $\{y_t\}$, $t \in T$, where T is usually equally divided. Time-series analysis aims to generate a time series of random processes and realize formulaic description by an appropriate statistical model[5].

The autoregressive(AR) model is a commonly used time-sequence model. Regression-based modeling is a complicated calculation process. Under the data grid environment, it is not reliable to use this method to accurately predict the node load over a long time. On the other hand, the prediction performance of the AR model is relatively stable and the calculation cost is low. Though the data grid environment is dynamic and complex, the simple AR model can achieve good short-term prediction. Therefore, the number of historical observations can be obtained by the regression model to obtain more accurate predicted values. That is, according to the order of the AR model, the number of historical observations can be obtained.

3.2 Support vector regression

According to the statistical theory, the support vector machine(SVM) learning method can be used to solve classification and regression problems. The SVM is used to solve the problem of support vector regression (SVR). When dealing with nonlinear problems, nonlinear problems are first transformed to linear problems in a high-dimensional space. Then, a kernel function is used to replace inner product operations in the high-dimensional space. Finally, complex computing problems are solved.

Regression can estimate the function between independent variables and dependent variables. According to this function, samples are input to obtain predicted values[6]. Suppose that $\{x_i, \ y_i\}_{i=1}^{l}$ is a training set containing l training samples, where the i-th input data $x_i \in \mathbf{R}^n$ and the i-th output data $y_i \in \mathbf{R}$. Here y_i can be any real number. A real-valued function $f(x)$ is founded to express the dependence of y on x. A linear regression problem should be transformed to an optimization problem[7]. By using the Lagrange function, the dual form of the original optimization problem can be obtained. By solving the dual problem, the linear regression function is finally obtained.

3.3 Design of SVR-based time-series prediction model

The SVR model can solve small-sample and high-dimensional problems. When the SVR model is used to predict data, an accurate SVR prediction model should be constructed. The key point is to select an appropriate kernel function and determine parameters to obtain high prediction accuracy[6].

The prediction model can be constructed by three steps as follows.

1) Sample acquisition and preprocessing

A model for the time series$\{x_1, x_2, \ldots, x_N\}$ is built and divided into two parts. The former n data are taken as a training set and used to construct the prediction model. The latter $N-n$ data are used for the prediction test.

Before the modeling prediction, all data should be normalized. That is, all data which have different indicators should be normalized to $[0, 1]$ or $[-1, 1]$ to reduce the calculation complexity in the training process and prevent the indicators which have the bigger values from controlling the training process.

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where x_i and x_i' are the sample data and the normalized data[8], respectively; $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the sample data, respectively.

When the input data of the model are normalized, the output data performs a denormalization operation. As a reverse process of normalization, denormalization is used to restore the calculation and obtain the actual value. The denormalization formula can be expressed as

$$x_i = x_i'(x_{\max} - x_{\min}) + x_{\min} \tag{3}$$

2) Model selection and parameter determination

In the SVR-based modeling, the kernel function is directly related to the performance of the established model. Each kernel function has its own applicable data distribution type. For different data sets, the performance of each kernel function is not the same. In the absence of the priori knowledge, the RBF kernel function is a better choice. The model trained by the RBF kernel function has better performance. The parameters have a significant impact on the SVR. With appropriate parameters, the SVR can have good learning and generalization ability.

3) Prediction evaluation method

There are a variety of methods and indicators on evaluating model prediction errors. A root mean square error (RMSE) is used to evaluate the performance of the prediction process, which can be calculated by

$$R = \sqrt{\frac{1}{M}\sum_{k=1}^{M}[y(k) - \hat{y}(k)]^2} \tag{4}$$

where $y(k)$ and $\hat{y}(k)$ are the measured and predicted values, respectively; M is the total number of samples.

Time-series prediction is another form of regression prediction. Regression prediction uses monitoring data at all times, where time is taken as an independent variable and the monitoring data are considered as dependent variables. By regression analysis, the function between the data and the time is established. The training process is to find the correlation function between the previous data and the following data.

3.4 ARSVR method

In the traditional time-series prediction methods, both the AR model and the SVR-based prediction model have their own advantages and disadvantages. The former presents a determinate model structure and provides a certain model order and a parameter estimation method. The latter does not require a complex statistical process, but directly trains the regression function based on the observational data and independently learns time-series variation. Combining the AR model and the SVR-based prediction model, a relatively simple and accurate model estimation and prediction method is obtained. The ARSVR load prediction method makes use of the AR model to model the sequence of historical observations. The model order and the dimensions of the vectors in the SVR are determined. The SVR is trained by using historical data, and a better regression function is eventually obtained to predict future loads.

3.4.1 Determination of AR model order and parameters

For a practical problem, when the random process or time observations $\{Y_t, t=0, 1, \ldots\}$ are related to or dependent on the previous observation Y_{t-1} or Y_{t-2}, the AR model can be used[9]. The linear equation of the AR model can be expressed as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \tag{5}$$

where $\phi_i (1 \leqslant i \leqslant p)$ is a real number and can be called an autoregressive parameter; ε_t is the residual, which is the white noise sequence with zero mean and σ^2 variance. Eq. (5) can be called the p-order autoregressive model, which is denoted as AR(p).

1) Order determination

The order of the AR model can be determined by the average information criterion (AIC), which is a criterion based on the judgement of the amount of information and can also be called the average information criterion[10]. The AIC formula can be expressed as

$$C(p) = N\ln\sigma_\alpha^2 + 2p \tag{6}$$

where $\sigma_\alpha^2 = S/(N-p)$, and S is the square of the residual; p is the model order; N is the number of data; σ_α^2 is the residual variance.

2) Parameter estimation

Parameter estimation runs in a given order. The model order cannot be determined in advance. During the modeling process, the model order should be set first. Then, the parameters of the AR model can be estimated by a parameter identification method and the model order can be obtained. Finally, the smallest order of $C(p)$ is taken as the optimal order and the AR model is determinate. After solving the AR model, the number of the input vectors of the SVR can be estimated.

3.4.2 Parameter determination of SVR

The SVR is a highly nonlinear system. Small changes of the internal parameters of the SVR can affect the system performance. The RBF kernel function is used as a kernel function. The parameters which need to be selected are the loss function parameter, the punishment parameter and the width coefficient.

1) Loss function parameter ε

The loss function parameter controls the number of support vectors. Its value is closely related to the sample noise. With the increase in ε, the number of support vectors decreases, which may cause the results that the model is too simple and the learning precision is not good enough. If ε is too low, the model may be too complex and cannot have a good generalization ability.

2) Punishment parameter P

The punishment parameter reflects the penalty level of the method when the sample data are beyond ε. Its value impacts the complexity and stability of the model. If P is too low, the penalty is small and the training error is great. With the increase in P, the learning accuracy increases and the generalization ability of the model becomes weak.

3) Width coefficient σ

The width coefficient reflects the correlation between support vectors. If σ is too low, the correlation becomes incompact and the learning machine is relatively complex. The generalization ability cannot be guaranteed. If σ is too large, the correlation is too strong and the regression model has difficulty in achieving sufficient accuracy.

4 Simulation Verification

In order to verify the validity of the above-mentioned prediction method, simulation experiments are carried out using Matlab 6.5. In these experiments, the Load Trace Playback Tool, which was developed by Dinda, is used to produce the background workloads of various machines and the actual situation of the CPU load is simulated[11]. During the simulation, the performance indicator of the host is simplified. The host load is replaced by the CPU utilization rate, and the frequency of the sampling is 1 Hz. 12 960 continuous history data are chosen as the original data, of which the former 11 000 data are the training data and the latter 1 960 data are used to test the model.

First, the AR model is solved to determine the model order. A search method is used and the highest order is set at 20. Then, the order of the estimated parameter is carried out. Finally, the AIC is used to judge the results.

During the modeling process, it can be found that no matter how high the order is, the accuracy of the model cannot be improved after the 16th band. So the 16th order model is selected. The former 11 000 history records are used to obtain the model AR(16). The input vector dimension of the SVR is randomly chosen as 4. The ARSVR is fixed and the parameters of the kernel function are dynamically adjusted. Then, the latter 1 960 data are tested. The error statistics is shown in Tab. 1.

Fig. 2 (a) is a change curve of the CPU utilization rate within a period of time. By using the ARSVR method, the regression function can predict the data in testing sets and a predicted curve is obtained, as shown in Fig. 2(b).

Tab. 1 Error statistics

Model	RMSE(200 previous data)	RMSE(1 960 previous data)
AR(16)	0.077 1	0.079 1
Normal SVR	0.069 3	0.070 0
Fixed ARSVR	0.052 9	0.053 1
Dynamic ARSVR	0.050 0	0.048 1

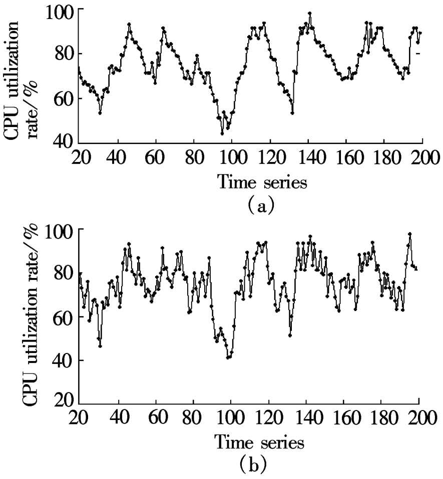

Fig. 2 Change curves of CPU utilization rate. (a) Experimental result; (b) Prediction result by using ARSVR method

From the error statistics, it can be seen that the dynamic ARSVR method has better prediction performance than that of the SVR and AR models. The dynamic ARSVR method can not only obtain the exact number of the history loads which works on the current loads by using the AR model, but also rapidly adjusts parameters to train a new model when the host load changes. By using the AR model, the model order can be obtained and the input vector dimensions of the SVR can be determined. Therefore, the ARSVR method has better performance than that of the SVR with randomly selected input vector dimensions.

5 Conclusion

Under a data grid environment, the operation characteristics of grid nodes are studied, and observational data are continuously analyzed. Then, the changing disciplines and pattern characteristics are obtained. During the prediction of the node load, only the load in a certain time is predicted by the current method, rather than the running task in the whole time. Except for having a strong relevance and self-similarity, the time series of the host load has a mutation in nature. Therefore, the predicted time series within a longer time may have a drastically changing behavior. The generation of these actions may be the derivation or end of the task from the host, or a new stage of implementation of the host. When predicting the host load, the load prediction of tasks in the entire running time should be studied to fully reflect the features of the host load.

References

[1] Foster I, Kishimoto H, Savva A, et al. The open grid services architecture version 1.5[EB/OL]. (2006-07-24)[2006-08-23]. http://forge.gridforum.org/projects/ogsa-wg.

[2] Ma Yongzheng, Nan Kai, Yan Baoping. MDS-based information service for data grid[J]. *Microelectronics &Comput-*

er, 2003, **20**(8): 27 − 30. (in Chinese)

[3] Dinda P, O'Halloran D. The statistical properties of host load [J]. *Scientific Programming*, 1999, **7**(3): 211 − 229.

[4] Huang Changqing, Xu Haishui, Zhong Huiyun. Host load prediction based on AR module[J]. *Computer Technology and Development*, 2007, **17**(9): 38 − 40. (in Chinese)

[5] He Shuyuan. *Using time series analysis*[M]. Beijing: Peking University Press, 2003: 234 − 238. (in Chinese)

[6] Ren Xunyi, Wang Ruchuan, Xie Yongjuan. Comparisons of SVM and LS-SVM for intrusion detection[J]. *Computer Science*, 2008, **35**(10): 83 − 85. (in Chinese)

[7] Zhang Haoran, Wang Xiaodong, Zhang Changjiang, et al. Learning algorithm for a new regression SVM[J]. *Journal of Test and Measurement Technology*, 2006, **20**(2): 168 − 173. (in Chinese)

[8] Sun Deshan, Wu Jinpei, Xiao Jianhua. The application of SVR to prediction of chaotic time series[J]. *Acta Simulata Systematica Sinica*, 2004, **16**(3): 519 − 521. (in Chinese)

[9] Xu Feng, Wang Zhifang. Research on AR model applied to forecast trend of vibration signals [J]. *Journal of Tsinghua University*: *Science and Technology*, 1999, **39**(4): 57 − 59. (in Chinese)

[10] Chen Shuncai. The research of time series prediction based on support vector machine[D]. Lanzhou: College of Electrical and Information Engineering of Lanzhou University of Technology, 2008. (in Chinese)

[11] Liu Renbin. *Matlab* 6 *computation engineering and application*[M]. Chongqing: Chongqing University Press, 2001. (in Chinese)

基于 ARSVR 方法的网格计算资源负载预测

黄　刚　王汝传　解永娟　石小娟

（南京邮电大学计算机学院，南京 210003）

摘要：在 MDS4 监控模型的基础上，设计了基于可靠存储与容侵数据网格的监控模型，分析了监控模型中计算资源的负载特性、指标．然后，设计了基于 SVR 的时间序列自回归预测模型，提出了用于数据网格负载预测的监控 ARSVR 方法．最后，利用 AR 模型对历史观测序列进行建模，确定模型的阶次．根据历史数据对 SVR 进行训练，得到回归函数．仿真实验结果表明，ARSVR 方法能对节点的负载进行有效预测．

关键词：网格；自回归支持向量回归机算法；计算资源；负载预测

中图分类号：TP393