Future Generation Computer Systems 81 (2018) 41–52

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Workload prediction in cloud using artificial neural network and adaptive differential evolution

Jitendra Kumar *, Ashutosh Kumar Singh

Department of Computer Applications, National Institute of Technology, Kurukshetra, India

HIGHLIGHTS

- The paper presents a workload prediction approach for cloud datacenters using neural network and self adaptive differential evolution.
- The proposed approach outperforms well known back propagation network approach in accuracy.
- The root mean squared error is used as accuracy measurement metric and proposed approach is able to achieve significant reduction in the prediction error.

ARTICLE INFO

Article history:
Received 10 January 2017
Received in revised form 26 July 2017
Accepted 22 October 2017

Keywords:
Cloud computing
Data center
Workload prediction
Neural network
Differential evolution

ABSTRACT

Cloud computing has drastically transformed the means of computing in recent years. In spite of numerous benefits, it suffers from some challenges too. Major challenges of cloud computing include dynamic resource scaling and power consumption. These factors lead a cloud system to become inefficient and costly. The workload prediction is one of the variables by which the efficiency and operational cost of a cloud can be improved. Accuracy is the key component in workload prediction and the existing approaches lag in producing 100% accurate results. The researchers are also putting their consistent efforts for its improvement. In this paper, we present a workload prediction model using neural network and self adaptive differential evolution algorithm. The model is capable of learning the best suitable mutation strategy along with optimal crossover rate. The experiments were performed on the benchmark data sets of NASA and Saskatchewan servers' HTTP traces for different prediction intervals. We compared the results with prediction model based on well known back propagation learning algorithm and received significant improvement. The proposed model attained a shift up to 168 times in the error reduction and prediction error is reduced up to 0.001.

1. Introduction

Cloud computing also referred as on-demand computing is expanded remarkably in the past few years. Companies are getting their applications up and running faster through the cloud. It has enabled users to store and process their data in third party data centers. Usage of virtual computing resources that are available as services over the Internet has gained significant attention from both industry and academia. In traditional computing, a user can have access to a fixed amount of computing resources. On the other hand, in cloud computing approach on demand resources are being offered to its customers [1]. Thus cloud has been very helpful in avoiding upfront infrastructure costs of companies. A wide number of organizations are switching to the cloud because of its characteristics like robustness, scalability, on demand service and several others. Service providers deploy huge data centers in order to provide on demand services to their customers. These data centers work as the backbone of the cloud computing environment. The key concept behind the cloud data centers is virtualization that helps in sharing resources among multiple users through virtual machines (VM). In order to maximize the gain of service providers along with maintaining the quality of services (QoS), data centers need an efficient and dynamic resource scaling and allocation policy.

The possibility of dynamic scaling along with advanced resource management is raised due to the flexibility of adding or removing virtual hardware resources at any point of time in the lifetime of hosted application [2]. Dynamic resource scaling has become a critical point of concern for a data center to work in a flawless manner. Resource scaling depends on several factors including number of active users, upcoming events, the current

* Corresponding author.
E-mail addresses: jitendrakumar@ieee.org (J. Kumar), ashutosh@nitkkr.ac.in (A.K. Singh).

https://doi.org/10.1016/j.future.2017.10.047
0167-739X/

state of the system and several others. The reactive scaling methods are not beneficial due to non instantaneous initialization and migration of virtual machines. While the proactive methods allow scaling of resources before actual demand arrives but these approaches require upcoming workload information in advance. The future resource demands can be predicted through identification of historical usage patterns and the current state of the data center. Since the predicted load on data center will determine the amount of resources to scale, it is required to set up an effective and reliable prediction system to achieve an accurate estimation of upcoming workload. This information can also be used to improve resource utilization and power consumption. Thus, the cloud data center can be operated at lower cost along with reducing SLA violations instances. The key challenges for precise predictions are interaction with varying number of clients and high non-linearity in workload. The workload can be predicted through several approaches. One can measure the maximum or average workload for specified time intervals. However, methods based on the statistics such as mean and maximum are very general and not capable of producing accurate predictions. For instance, if we use a prediction model based on maximum workload then resources will remain unused most of the time. On the other hand, if the average approach is used to predict future workload then the system will witness the lack of resources. This will cause performance degradation whenever workload increases. The prediction models based on such methods are considered to be poor as they are not good in prediction and correspond to a very few number of cases [3]. Machine learning methods are widely being adopted for establishing more accurate prediction models. Machine learning techniques use historical data as training window to predict the workload throughout a prediction interval [4]. Where prediction interval defines the time between each prediction.

This paper contributes towards the development of workload prediction model based on neural network and self adaptive differential evolution that can predict the workloads with higher accuracy. Instead of using simple statistics such as mean and others, the proposed model learns and extract the pattern from workload. The obtained patterns are used for further predictions. The model is trained using evolutionary approach to minimize the effect of initial solution choice. The evolutionary algorithm explores the space in multiple directions using a set of solutions. One of the difficult tasks in evolutionary algorithms is the parameter tuning. This effect is minimized as the proposed predictive model is capable in learning the crossover rate, mutation rate and mutation strategy too. These predictions can be further utilized in improving the resource scaling decisions. The model has been tested on the benchmark data sets of two servers and compared with the back propagation neural network model. The proposed model outperformed the prediction models based on average, maximum and back propagation network with a substantial reduction in mean squared prediction error.

Rest of the paper is organized as follows: Section 2 provides an overview of related work. The proposed approach is discussed in Section 3 followed by results and discussion in Section 4. Finally, the paper is wrapped up with conclusive remarks and future scope in Section 5.

2. Related work

Many of the researchers are working in the area of workload prediction and they have addressed the problem with the help of different approaches. Two well known approaches are homeostatic prediction and history-based prediction. In homeostatic prediction, the upcoming workload at next time instance is predicted by adding or subtracting a value such as the mean of previous workloads from the current workload [5]. The value to be subtracted or added can be static (a fixed number) or dynamic (estimated value from previous workload instances). On the other hand, history based methods are simple and popular prediction models. These models analyze previous workload instances and extract patterns to predict the future demands. History based methods use tendency information of previous workload instances while homeostatic model goes back to the mean of previous workloads [6].

In order to use historical information, data center workloads exhibiting seasonal trends can be presented in the form of time series [7]. A set of data points measured at successive points in time spaced at uniform time intervals is called time series data. Classical methods have been used extensively in time series prediction. This group of models includes Exponential Smoothing (ES), auto regression (AR) model, Moving Average (MA) model, Autoregressive Integrated Moving Average (ARIMA) model, Hidden Markov Model (HMM) and many others.

In [8], the authors have proposed an auto regression based method to predict the web server workload but the model is strictly linear in nature. In auto regression, a linear combination of past values of the variable under consideration is used to forecast the value for upcoming time instances. Danilo et al. [9] used moving average to develop a prediction model. A distributed solution was proposed that incorporated workload prediction and distributed non linear optimization techniques. In [10], Kalekar used exponential smoothing for seasonal time series prediction. Two different approaches multiplicative seasonal model and additive seasonal model were used to predict the workload. The model is appropriate for time series exhibiting seasonal behavior only. Roy et al. [3] proposed a forecasting model using Auto Regressive Integrated Moving Average (ARIMA) model. The authors also discussed the challenges involved in auto scaling in a cloud environment. In [11], the authors have used regression techniques to analyze live sports event broadcast service workloads from a commercial Internet service provider. The approach is based on simple statistical models that might not capture the patterns in more complex data. Khan et al. [12] presented a workload prediction model based on multiple time series approach. The model does a grouping of similar applications' need in order to improve the accuracy of predictions. The authors also utilized hidden Markov model (HMM) to distinguish the temporal correlations in obtained clusters of VMs. They use this information to define the variations in workload patterns over time. Other methods also have been used in workload forecasting such as Monte Carlo [13].

Apart from classical approaches machine learning also has been widely explored for time series prediction. Machine learning methods can learn from data provided to them and give some probabilistic score on an unknown pattern from its past experience. In [14] authors proposed a framework using self organizing map (SOM) and support vector machines (SVMs). Self organizing map was used to cluster the data in different regions while SVMs were used to predict the future data. But the approach is highly sensitive to the threshold value that decides the number of data points in a partitioned region. A two tier architecture using k Nearest Neighbors (kNN) was proposed by Tao Ban et al. for financial time series prediction [15] but $kNNs$ are lazy learners and need high computational cost. The Neural network was used in [16] to model workload variations in multimedia designs. In [17] authors presented a resource scaling method based on workload prediction. Linear regression was used for predicting workload. The predicted workload was used to decide the type of scaling.

Hu et al. [18] used the statistical learning theory to build a prediction model that integrated with support vector regression (SVR) and Kalman smoother. It achieved high prediction accuracy compared to the auto regression, back propagation network and standard SVR. The predictions were used for further resource scaling decisions. In [19] authors implemented an ensemble based