1-1-2016

RVLBPNN: A Workload Forecasting Model for Smart Cloud Computing

Yao Lu
Jiangsu University

John Panneerselvam
University of Derby

Lu Liu
Jiangsu University

Yan Wu
Boise State University

Publication Information

Lu, Yao; Panneerselvam, John; Liu, Lu; and Wu, Yan. (2016). "RVLBPNN: A Workload Forecasting Model for Smart Cloud Computing". *Scientific Programming, 2016*, 5635673-1 - 5635673-9. http://dx.doi.org/10.1155/2016/5635673

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5635673, 9 pages
http://dx.doi.org/10.1155/2016/5635673

Research Article

RVLBPNN: A Workload Forecasting Model for Smart Cloud Computing

Yao Lu,[1] John Panneerselvam,[2] Lu Liu,[1,2] and Yan Wu[1,3]

[1]*School of Computer Science and Telecommunication Engineering Jiangsu University, Jiangsu, China*
[2]*Department of Computing and Mathematics, University of Derby, Derby, UK*
[3]*Department of Computer Science, Boise State University, Boise, USA*

Correspondence should be addressed to Lu Liu; l.liu@derby.ac.uk

Received 28 July 2016; Accepted 19 September 2016

Academic Editor: Wenbing Zhao

Given the increasing deployments of Cloud datacentres and the excessive usage of server resources, their associated energy and environmental implications are also increasing at an alarming rate. Cloud service providers are under immense pressure to significantly reduce both such implications for promoting green computing. Maintaining the desired level of Quality of Service (QoS) without violating the Service Level Agreement (SLA), whilst attempting to reduce the usage of the datacentre resources is an obvious challenge for the Cloud service providers. Scaling the level of active server resources in accordance with the predicted incoming workloads is one possible way of reducing the undesirable energy consumption of the active resources without affecting the performance quality. To this end, this paper analyzes the dynamic characteristics of the Cloud workloads and defines a hierarchy for the latency sensitivity levels of the Cloud workloads. Further, a novel workload prediction model for energy efficient Cloud Computing is proposed, named RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network) based on BPNN (Backpropagation Neural Network) algorithm. Experiments evaluating the prediction accuracy of the proposed prediction model demonstrate that RVLBPNN achieves an improved prediction accuracy compared to the HMM and Naïve Bayes Classifier models by a considerable margin.

1. Introduction

Cloud Computing is emerging as a prominent computing paradigm for various business needs, as it is known to be a low-cost any-time computing solution. The on-demand service access features of the Cloud Computing help the Cloud clients to adopt or transform their business model to Cloud datacentres for computing and storage resources [1]. This increasing number of Cloud adoptions by various business domains over the recent years is also reflected in the increase in the number of Cloud service providers. An immediate impact of this is that Cloud datacentres are addressed to be one of the major sources of energy consumers [2] and environmental pollutants. To this end, Cloud datacentres are addressed to be causing energy, economic, and environmental impacts to an irresistible margin. It has been reported [3] that ICT (Information Communication Technology) energy consumption will contribute up to 50% of the total energy expenditures in the United States in the next decade, which was just 8% in the last decade. Energy efficient computing has been promoted and researched under various dimensions for the purpose of reducing the energy consumption levels of the datacentre whilst processing workloads and cooling the server resources. It is worthy of note that cooling system in a typical Cloud datacentre would incur considerable amount of energy cost of those spent towards the actual task execution [4]. Thus it is apparent that energy efficient Cloud Computing is one of demanding characteristics of Cloud Computing.

Resource management driven by forecasting the future workloads is one of the possible ways of achieving energy efficiency in Cloud Computing. In general, the intrinsic dynamic [5] nature of the Cloud workloads imposes complexities in scheduling, resource allocation, and executing workloads in the datacentres. Predicting the nature of the future workloads

can help reduce the energy consumption levels of the server resources by the way of effectively scheduling the incoming workloads with the most appropriate level of resource allocation. Alongside energy efficient computing, predictive analytics in Cloud Computing also benefits [5] effective resource utilization, optimum scalability of resources, avoiding process failures, capacity planning, network allocation, task scheduling, load balancing, performance optimization and maintaining the predetermined QoS (Quality of Service) and SLAs (Service Level Agreements), and so forth.

Owing to the extravagant dynamicity of Cloud workloads, understanding the characteristic behaviors of the Cloud workloads at the datacentre environment is often a complex process. Mostly, Cloud workloads are of shorter duration and arrive more frequently at the datacentres and are generally not computationally more intensive unlike scientific workloads. Furthermore, every submitted workloads are bound to a certain level of latency sensitiveness [6] which decides the time within which the workload has to be processed. Workloads with increased latency sensitivity levels usually demand quicker scheduling from the providers. This implies that an effective prediction model should possess the qualities of understanding the inherent characteristics and nature of the Cloud workloads and their corresponding behaviors at the datacentres.

Despite the existing and ongoing researches, Cloud Computing still demands extensive analyses of the Cloud entities for the purpose of modelling the relationship between the users and their workload submissions and the associated resource requirements. An effective prediction model should necessarily incorporate the knowledge of three important characteristic events in a datacentre environment in order to achieve reliable level of prediction accuracy. Firstly, the volume and the nature of the workloads submitted are driven by the users based on their requirements and resource demands. Increased amounts of jobs submissions obviously demand increased amounts of resource allocation and thus causes increased energy expenditures. Secondly, the actual execution of the workloads would not necessarily consume all the allocated resources. The immediate implication is that increased proportions of allocated resources remain idle during task execution and incur undesirable energy consumptions. Finally, the user behavioral pattern of job submission and associated resource consumption are subjected to change over time.

The intrinsic dynamism of both the Cloud workloads and the server resources should be effectively captured [7] by the prediction model over a prolonged observation period. Existing works in analyzing the intrinsic characteristics of the Cloud entities have not contributed suffice inferences [5, 8, 9] required for an effective prediction model. Imprecise knowledge of such aforementioned parameters of the Cloud entities would increase the prediction error margin, which would directly affect the Quality of Service (QoS) by violating the Service Level Agreement (SLA). With this in mind, this paper proposes a novel forecast model named RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network), based on an improved BPNN (BP Neural Network) for accurately predicting the user requests. Exploiting the latency sensitivity levels of the Cloud workloads, our proposed model predicts the user requests anticipated in the near future in a large-scale datacentre environment.

The rest of this paper is organized as follows: Section 2 introduces the previous works in Cloud workload prediction modelling. Section 3 presents a background study on Cloud workloads, exhibiting the dynamic nature of the Cloud workloads. The computational latencies affecting the Cloud workloads are defined in Section 4 and Section 5 proposes our prediction model based on the modified BP Neural Network. Our experiments are presented in Section 6 and Section 7 concludes this paper along with our future research directions.

2. Related Works

A number of researches are being conducted with the motivation of promoting green computing in the recent past. For instance, the approach of capacity management and VM (Virtual Machine) placement have been the strategies of [10, 11]. A workload placement scheme, called BADP, combines task's behavior to place data for improving locality at the cache line level. Further, [11] proposes a remaining utilization-aware (RUA) algorithm for VM placement. In general, workload placement and task allocation can be more effective when driven by a proactive prediction of the incoming workloads. Time series [12] approach incorporates the repeatable behaviors such as periodicity and timely effects of the various Cloud entities such as VMs and users and explores the temporal and spatial correlations in order to provide the prediction results. However, such technique usually explores the entities individually and often leads to inaccurate results resulting from the random behaviors of the individual entities.

A multiple time series approach [13] has been proposed to improve the prediction accuracy, by the way of analyzing the Cloud entities at the group level rather individually. Nonlinear time series approach works with the assumption that the observations are real valued and such techniques often require special emphasis on extracting the chaotic invariants for prediction analysis. Autoregression (AR) is a prediction technique [14] which usually predicts the next state transition by recursively acting on the prediction values. However, AR method has a conspicuous shortcoming that the prediction errors will be accumulated for long term prediction analysis because of the recursive effect. Another drawback of AR methods is that they only deliver accurate forecasts for datasets characterized with reasonable periodicity, which is shown in [15], where a number of different linear prediction models based on AR have been deeply analyzed. Poisson process [16] models the incoming workload arrival pattern for prediction analysis and has the capability of capturing complex nonexponential time varying workload features. Moving average approaches [14, 16] such as first-order and second-order moving average techniques used for prediction analysis cannot capture important features required to adapt to the load dynamics.

Recently, Bayes and Hidden Markov Modelling (HMM) [5] approaches were analyzed in our earlier works for evaluating their prediction efficiency in Cloud environments.

Byes technology predicts the future samples based on a predefined evidence window. The adjacent samples contained in the evidence window should be mutually correlated for delivering a reliable prediction output. Thus Bayes model will lose efficiency in a dynamic Cloud environment. However, Bayes model could still be deployed in situations where there are less fluctuations among the workload behavior. HMM is a probabilistic approach which is used to predict the future sate transition from the current state. In spite of the dynamic nature of the Cloud workloads, probabilistic approach may not scale well for predicting the future workloads with a reliable level of prediction accuracy. Despite the existing works, Cloud Computing still demands a smart prediction model with the qualities of relative high precision and the capacity of delivering a reliable level of prediction accuracy. With this in mind, this paper proposes a novel prediction model named RVLBPNN. Exploiting the workload characteristics, our proposed model achieves a reliable level of prediction accuracy. Our proposed model has been sampled and tested for accuracy based on a real life Cloud workload behaviors.

3. Background

3.1. Cloud Workloads. Cloud workloads arrive at Cloud datacentres in the form of jobs [15] submitted by the users. Every job includes certain self-defining attributes such as the submission time, user identity, and its corresponding resource requirements in terms of CPU and memory. A single job may contain one or more tasks, which are scheduled for processing at the Cloud servers. A single task may have one or more process requirements. Tasks belonging to a single job may also be scheduled to different machines but it is desirable to run multiple processes of a single task in a single machine. Tasks are also bound to have varied service requirements such as throughput, latency, and jitter, though they belong to the same job. The tasks belonging to the same job not necessarily exhibit higher correlating properties among them. Thus, tasks within the same job might exhibit greater variation in their resource requirements. Tasks might also interact among each other during their execution. Furthermore, two jobs with the same resource requirements may not be similar in their actual resource utilization levels because of the variation found among the tasks contained within the jobs. Based on the resource requirements, tasks are scheduled either within the same or across different servers. Usually, the provider records the resource utilization levels of every scheduled task and maintains the user profiles.

The attributes encompassed by the Cloud workloads, such as type, resource requirements, security requirements, hardware, and network constraints, can be exploited to derive the behaviors the Cloud workloads. Interestingly, Cloud workloads behave distinctively with different server architectures. Such distinctive workload behaviors with different server architectures strongly influence the CPU utilization, with the memory utilization generally remaining stable across most of the server architectures. Thus the resource utilization highly varies across the CPU cores compared to the memory or disc, as the disc utilization mostly shows similar utilization patterns across different server architectures. Thus the behaviors of workloads at the Cloud processing environment are strongly correlated with the CPU cores compared to RAM capacity of the machines at the server level. The capacity levels of CPU and memory in a physical server usually remain static. Resource utilization levels are more dynamic and vary abruptly under different workloads. Such dynamic parameters of the server architectures are usually calculated as the measure of the number of cycles per instruction for CPU and memory access per instruction for memory utilizations, respectively. Thus the task resource usage is usually expressed as a multidimensional representation [5] encompassing task duration in seconds, CPU usage in cores, and memory usage in gigabytes. It is commonly witnessed that most of the allocated CPU and memory resource are left unutilized during task execution.

3.2. Characterizing Workloads. User demand often changes over time which reflects the timely variations of the resource consumption levels of the workloads generated as they are driven by the users. User demands are generally influenced by the time-of-the-day effects, showing a repeating pattern [13] in accordance to the changing business behaviors of the day and by the popular weekend effects showing weekend declines and weekday increase trend in the arrival of the workloads. The relationships [17] between the workloads and user behaviors are primarily the integral component in the understanding of the Cloud-based workloads and their associated energy consumptions.

Different workloads will have different processing requirements such as CPU, memory, throughput, and execution time, and this variation results from the characteristic behaviors of different users. Nowadays, Cloud environments are more heterogeneous composing different servers with different processing capacities. In order to satisfy the diverse operational requirements of the Cloud user demands, normalization of this machine heterogeneity is now becoming an integral requirement of the Cloud providers, by which virtually homogenizing the heterogeneous server architectures and thereby eliminating the differentiation found in both the hardware and the software resources. In general, the different forms of workloads from the provider's perspectives include computation intensive with larger processing and smaller storage, memory intensive with larger storage and smaller processing, workloads requiring both larger processing and larger storage, and communication intensive with more bandwidth requirements. Workloads are usually measured in terms of the user demands, computational load on the servers, bandwidth consumption (communication jobs), and the amount of storage data (memory jobs). User demand prediction modelling requires an in depth quantitative and qualitative analysis of the statistical properties and behavioral characteristics of the workloads including job length, job submission frequency, and resource consumption of the jobs, which insists that the initial characterization of the workloads is more crucial in developing an efficient prediction model. Rather than the stand-alone analysis of the above stated workload metrics, modelling the relationships between them across a set of workloads is more significant in order to achieve more reliable prediction results. Statistical

properties [18] of the workloads are more significant for the prediction accuracy since they remain consistent in longer time frames. Some of the important characteristics of Cloud workloads affecting prediction accuracy include job length, job submission frequency, resource request levels, job resource utilization, and self-similarity.

3.3. Categorizing Workloads. In the Cloud Computing service concept, the workload pattern, the Cloud deployment types (public, private, hybrid, and community), and the Cloud service offering models (SaaS, PaaS, and IaaS) are closely interconnected with each other. From the perspectives of the Cloud service providers, the incoming Cloud workloads can be categorized into five major types [9] as static, periodic, unpredictable, continuously changing, and once-in-a-lifetime workloads. Static and periodic workloads usually follow a predictable pattern in their arrival frequency. Continuously changing workloads exhibit a pattern of definite variations characterized by regular increasing and declining trend in their arrival frequencies. Unpredictable workloads exhibit a random behavior in their arrival frequency and are the most challenging type of workloads for prediction analysis. Once-in-a-lifetime workloads are the rarely arriving workloads and their submissions are mostly notified by the clients.

4. Workload Latency

Latency plays an important role at various levels of processing the workloads in a Cloud processing infrastructure. This paper mainly focuses the influence of the workload latency sensitivity upon prediction accuracy. The most dominating types of latencies are the network latency and the dispatching latency, both of which actually result from the geographical distribution of the users and the Cloud datacentres. Both of these latencies depend on the Round Trip Time (RTT) [19], which defines the time interval between the user requests and the arrival of the corresponding response. Another type of latency existing in the process architecture is the computational latency which is the intracloud latency [20] found among the processing VMs located within a single datacentre. This latency depends on both the software and the hardware components [6, 21] such as CPU architecture, runtime environment, and memory, guests and host operating system, instruction set, and hypervisor used. CPU architecture, Operating System, and the scheduling mechanisms are the most dominating factors of this type of in-house computing latency, and efficient handling of such resources helps reducing the impacts of the computational latencies.

Jobs submitted at the Cloud datacentre undergo various levels of latencies depending on the nature of their process requirements and the end-user QoS expectations. Since a single job might contain a number of tasks, the latency sensitivity of every single tasks has to be treated uniquely. A single definition of the computing latency cannot fit all types of jobs or tasks, since every job is uniquely viewed at the datacentre. For instance, processing a massive scientific workload may span across several days or months, in which latencies of a few seconds are usually acceptable. Common example of the latency sensitive workloads is the World Wide Web, among which different applications have different latency levels. The acceptable level of latencies is usually the measure of the end-user tolerances. Workloads resulting from users surfing the internet are generally latency-insensitive. Jobs including online gaming and stock exchange data are the commonly witnessed latency sensitive applications. The level of sensitivity is determined by the allowed time-scale for the providers to provide an undisrupted execution of the workloads for delivering the desired levels of QoS, ranging from a few microseconds to a few tens of microsecond end-to-end latencies.

The taxonomy of the latency levels of the Cloud workloads studied in this paper are attributed from level 0 representing the least latency sensitive tasks to level 3 representing the most latency sensitive tasks. Least latency sensitive tasks (level 0) are nonproduction tasks [22] such as development and nonbusiness critical analyses, which do not have a significant impact on the QoS even if these jobs are queued at the back end servers. Level 1 tasks are the next level of latency sensitive tasks and are generally the machine interactive workloads. Level 2 tasks are the real time machine interactive workloads and the latency tolerance levels of these tasks stay at tens of milliseconds. Level 3 tasks are the most latency sensitive tasks with latency tolerance levels at the range of submilliseconds, and are generally the revenue generating user requests such as stock and financial analysis. Workloads characterizing an increased level of latency sensitivity are usually treated with higher scheduling priorities at the datacentres. Latency analysis has a prime importance in greening the datacentre, since every job or task submitted to the Cloud has its own level of latency tolerances, directly affecting not only the various workload behaviors at the datacentres but also the end-user QoS satisfaction.

Based on our earlier analysis conducted on a Cloud dataset [23], we perform a latency aware quantification of the jobs submitted to the datacentre comprising a total of 46093201 tasks in our recent work [24]. Figure 1 illustrates a day-wise submission of tasks across the observed 28 days. Figure 2 quantifies the total number of task submissions in terms of their latency sensitivity levels. It can be observed that most of the task submissions are least latency sensitive accounting for 79.52% of the total task submissions, followed by level 1, level 2, and level 3 with 12.46%, 7.54%, and 0.47%, respectively.

5. Proposed Prediction Model

This section describes our novel prediction model aimed at predicting the anticipated workloads in a large-scale datacentre environment.

5.1. BP Neural Network Method. BP Neural Network is a multilayer hierarchical network composed of upper neurons and fully associated lower neurons. Upon training the input samples into this multilayer network structure, the transformed input values are propagated from the input layer through the middle layer, and the values are outputted by the neurons in the output layer. The error margins

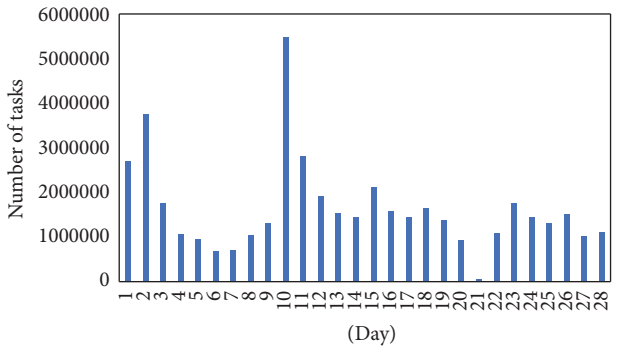

Figure 1: Total number of task submissions.

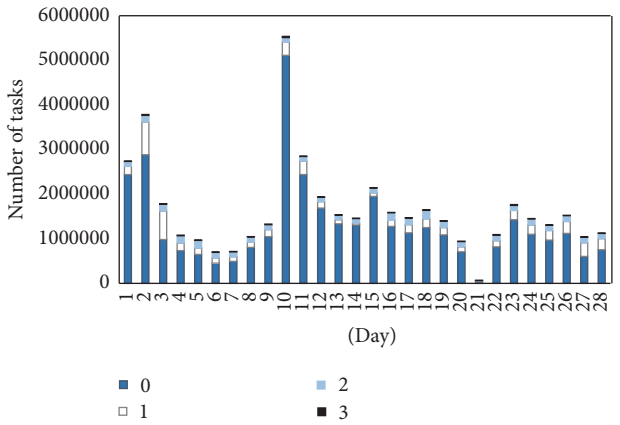

Figure 2: Latency-wise task submission.

between the actual and the expected output are normalized by the way of the output neurons adjusting the connection weights of the neurons in both the middle layer and the input layer. This back propagation mechanism of connection weight adjustment enhances the correctness of the network responses of the neurons to the input values. As the BP algorithm implements a middle hidden layer with associated learning rules, the network neurons can effectively identify the hidden nonlinear pattern among the input samples.

5.2. BP Neural Network Architecture. A typical neuron model can be derived according to characteristics of the neurons [25–28], which is shown in Figure 3. In this figure, $X_1, X_2, \ldots, X_n$ are n input data to the neurons; $a_{i1}, a_{i2}, \ldots, a_{in}$ are the weight factor of $X_1, X_2, \ldots, X_n$, respectively; $g()$ is a nonlinear function; O_i is the output result; and λ_i is the threshold.

Based on the above neuron structure, we make $O_i = g(P_i)$, where, $P_i = \sum_{j=1}^{n} a_{ij}X_j - \lambda_i$. In the formula, X represents the input vector, a_i represents the connection weight vector for neuron i, and P_i is the input of the neurons. In most cases, λ_i is considered to be the 0th input of the neuron. Thus, we can get a simplified equation of the above expression, which is shown in formula (1). In this equation, value $X_0 = -1$, and $a_{i0} = \lambda_i$.

$$P_i = \sum_{j=0}^{n} a_{ij}X_j. \quad (1)$$

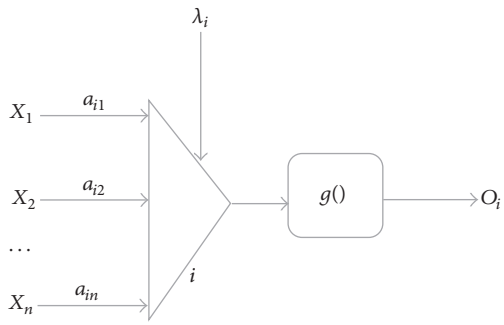

Figure 3: Neuron model.

5.3. An Improved BP Neural Network Algorithm for Prediction. BPNN can effectively extract the hidden nonlinear relationships among the Cloud workloads. However, BPNN with a fixed learning rate cannot extract this nonlinear relationships among the samples of large datasets, since BPNN has a slow convergence rate for large-scale datasets in the range of Big Data. A modified BP algorithm named VLBP (variable learning rate backpropagation) has been proposed to enhance this convergence rate [29]. In comparison with the BP algorithm, VLBP has a characteristic enhancement in both the computation speed and precision of the output. But the VLBP algorithm can be susceptible to several numbers of local minima resulting from the irregular shake surface error. This slows down the update process of Mean Square Error (MSE) and increases the presence of local minimum points. This results in a higher approximation precision despite the improvement in the convergence rate. VLBP exhibits a fluctuating and slower learning process and increases the length of the computation.

This necessitates further improvements in the BP Neural Networks for the purpose of enhancing its prediction efficiency whilst training large datasets. This paper proposes a novel prediction method using a modified BP algorithm by incorporating variant conceptions of a genetic algorithm. Our proposed prediction method effectively adjusts the learning rate of the neurons to a certain probability in accordance with the trend of the MSE during the execution of the VLBP algorithm. The learning rate may not be changed or multiplied by the factor ρ greater than 1 when MSE increases beyond the set threshold ζ.

Our proposed prediction algorithm is described as follows:

(1) Generate a random number rand(u) ($0 <$ rand(u) < 1).

(2) If rand(u) is less than a defined value (set as 0.8 in experiments), then execute VLBP algorithm.

(3) If rand(u) is greater than the defined value, else if MSE has increased, then the learning rate is multiplied by a factor greater than 1 despite MSE exceeding ζ or not; if MSE decreases after updating the connection weight, then the learning rate is multiplied by a factor between 0 and 1.

We named our proposed algorithm as RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network).

Through this method, the learning rate of the neurons will not be decreased at any time resulting from the slow renewal of MSE near the local minimum point. But, there is also a certain probability of increasing the learning rate of the neurons. RVLBP algorithm can identify the global minimum point by effectively avoiding the local minimum points. Thus our proposed algorithm reduces the presence of local minimum points during the learning process, thereby improving the learning efficiencies of the network neurons.

6. Performance Evaluation

6.1. Experiment Sample. This section demonstrates the efficiency of our proposed prediction model based on RVBLPNN. We train the input data sample to predict the anticipated values in the near future representing the future workloads expected to arrive at the datacentre. The experiment samples are trained in MATLAB 7.14 and the test datasets used are the publically available Google workload traces [23]. The datasets are a collection of 28 days of Google usage data workloads consisting of 46093201 tasks comprising CPU intensive workloads, memory-intensive workloads, and both CPU and memory-intensive workloads. The dataset parameters include time, job id, parent id, number of cores (CPU workloads), and memory tasks (memory workloads), respectively, to define the sample attributes. We compare the prediction efficiencies of our proposed prediction model against the efficiencies of Hidden Markov Model (HMM) and Naïve Bayes Classifier (NBC); both of them were evaluated in our earlier works [5]. All the three models are evaluated for their efficiencies in predicting memory and CPU intensive workloads accordingly. We train the prediction model with a set of 10 samples and contrast the prediction output with the actual set of successive 10 samples.

MATLAB simulation environment provides a built-in model for RVLBPNN technique, modelling RVLBPNN as a supervised learning. The Neural Network is comprised of a three-layer network structure. This three-layer Neural Network can approximate any type of nonlinear continuous function in theory. We ultimately use 10 input nodes, 12 hidden nodes, and 10 output nodes through a number of iterations for enhancing the prediction accuracy. The data samples are normalized and imploded in the interval (0, 1). "logsig" function is selected as the activation function of input layer, hidden layer, and the output layer, so that the algorithm exhibits a good convergence rate. Further, variable learning rate and random variable learning rate are adopted, respectively. This experiment uses 100,000 workload data samples as the training data and another 100000 data samples as the reference data. The prediction accuracy is the measure of correlations between the predicted and actual set of sample values.

6.2. Result Analysis and Performance Evaluation

6.2.1. Memory Workloads Estimation. Figure 4 depicts the estimation results of RVLBPNN, HMM, and NBC model, respectively, in terms of their prediction accuracy whilst predicting the memory-intensive workloads. The number of test

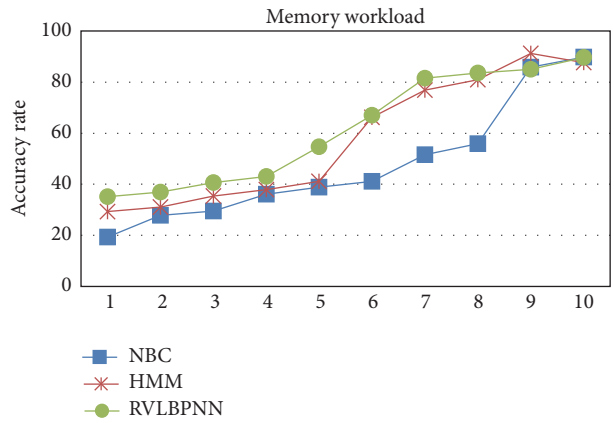

Figure 4: Prediction of memory-intensive workloads.

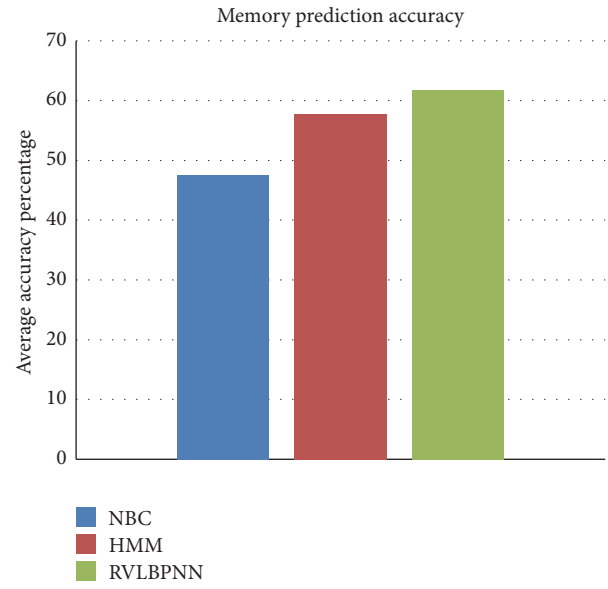

Figure 5: Prediction accuracy for memory-intensive workloads.

samples (x-axis) are plotted against the prediction accuracy (y-axis) for the three models; every set of sample consists of 10000 workload samples. For presenting the test results with a better interpretation, the sample results are sorted ascendingly from 1 to 10 based on the prediction results. The average accuracy percentage in estimating the memory-intensive workloads without considering the latency levels of individual workloads for NBC, HMM, and RVLBPNN are 47.69%, 57.77%, and 61.71%, respectively, as shown in Figures 4 and 5. It is evident from Figures 4 and 5 that the RVLBPNN exhibits a better prediction accuracy than both HMM and NBC techniques. It can be depicted from the estimation results that our proposed RVLBPNN model is demonstrating a minimum of 3% prediction accuracy better than HMM and 13% better than NBC, respectively.

This improved prediction accuracy of the RVLBPNN model is attributed to its ability of capturing the intrinsic relationship features among the arriving Cloud workloads. We further evaluate the efficiency of our proposed model in

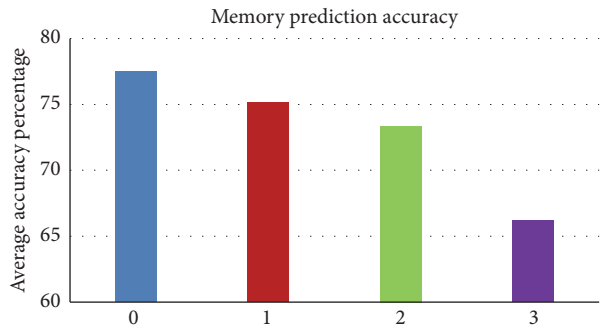

Figure 6: Latency-wise prediction accuracy for memory workloads.

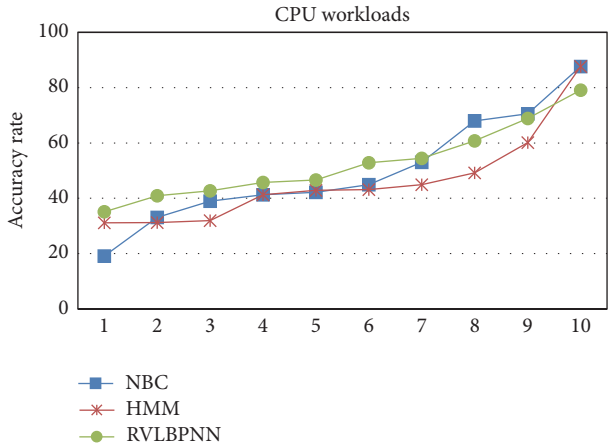

Figure 7: Prediction of CPU intensive workloads.

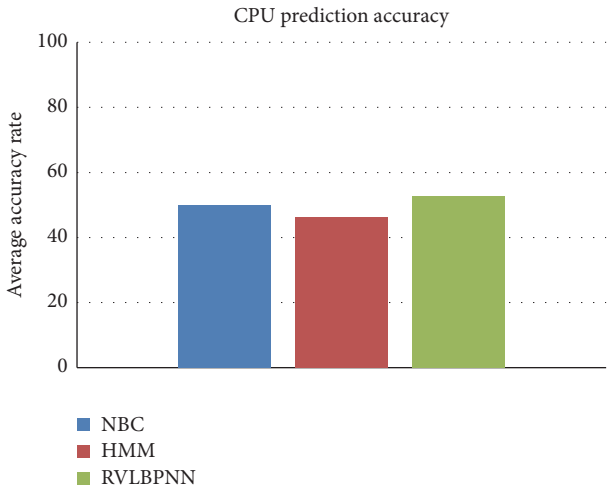

Figure 8: Prediction accuracy for CPU intensive workloads.

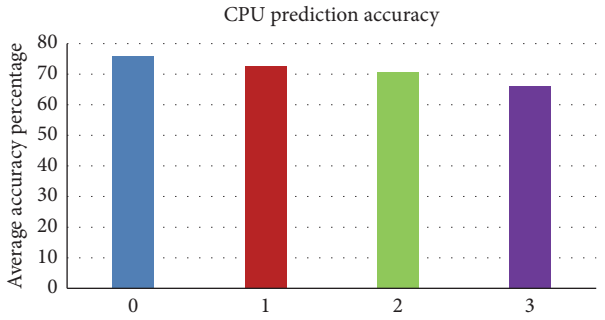

Figure 9: Latency-wise prediction accuracy for CPU workloads.

forecasting memory-intensive workloads of different latency sensitivity levels. Figure 6 depicts the estimation results of our proposed RVLBPNN model in terms of their prediction accuracy whilst predicting memory-intensive workloads of different latency sensitivity levels as described earlier in Section 4. It can be observed from Figure 6 that less latency sensitive memory workloads exhibit better predictability, with the prediction accuracy being 66.17% for level 3 workloads and 77.48% for level 0 workloads, respectively.

6.2.2. CPU Workloads Prediction. Similar to the memory-intensive workloads, the experiments are repeated for the CPU intensive workloads from the dataset. Figure 7 depicts the estimation results of RVLBPNN, HMM, and NBC whilst predicting the CPU intensive workloads. The average prediction accuracy of NBC, HMM, and RVLBPNN models is 49.87%, 46.36%, and 52.70%, respectively, whilst predicting CPU intensive workloads, as shown in Figure 8. It can be observed that RVLBPNN exhibits better prediction accuracy than both HMM and NBC models by a margin of around 3% and 6%, respectively.

We further evaluated the efficiency of our proposed prediction model in predicting the CPU intensive workloads of different latency levels. Figure 9 depicts the estimation results of our proposed RVLBPNN model whilst predicting the CPU intensive workloads of different latency sensitivity levels. We observe a similar trend of prediction accuracy between both memory and CPU workloads of different latency sensitivity levels. Again CPU intensive workloads of less latency levels are exhibiting better predictability, with the accuracy being 66.08% for level 3 workloads and 75.90% for level 0 workloads. This leads us to infer that least latency sensitivity level workloads exhibit a better rate of prediction accuracy for both CPU and memory-intensive workloads.

6.2.3. Interpretation and Discussion. From the experiment results, it is clearly evident that our proposed RVLBPNN model demonstrates better prediction accuracy than both HMM and NBC models by a considerable margin. Our proposed model outperforms the other two models whilst predicting both the CPU intensive and memory-intensive workloads. Meanwhile, we also observe that increasing levels of latency sensitivity of both CPU and memory-intensive workloads impose increasing error margin in the prediction results. Lower level of latency sensitivity exhibits better predictability. Since the majority of the Cloud workloads are of lower latency sensitivity levels, our proposed prediction model can accurately predict the trend of most of the arriving workloads. An increased level of intrinsic similarity among the arriving workloads facilitates a better learning rate of the neurons in the RVLBPNN model, which results in an increased prediction accuracy. From the experiments, we postulate that workloads should be treaded uniquely with

respect to their computational demand latency sensitivity and user requirements for achieving a reliable level of prediction accuracy. Furthermore, workload prediction analytics can be benefitted with better accuracy when the workloads are analyzed at the task level rather than at the job level.

7. Conclusion

Green computing has turned out to be one of the important characteristics for achieving sustainable smart world in the future. Resource management by the way of predicting the expected workloads facilitates optimum scaling of the server resources, reducing the presence of idle resources and allocating appropriate levels of server resources to execute the user requests. The reliability and accuracy levels of such prediction techniques directly impacts important decision making in large-scale Cloud datacentre environments. In this paper, we propose a novel workload prediction model for the purpose of predicting the future workloads in Cloud datacentres. Our proposed novel workload prediction model, called RVLBPNN, is based on BP Neural Network algorithm and predicts the future workloads by the way of exploiting the intrinsic relationships among the arriving workloads. The experimental results indicate that the proposed RVLBPNN model achieves better precision and efficiency than the HMM-based and NBC-based prediction techniques. As a future work, we plan to explore the possibilities of further improving the prediction accuracy of our proposed approach. For instance, incorporating the periodicity effects of the workload behavior into RVLBPNN can further enhance the prediction accuracy. Meanwhile, investigating the efficiencies of our novel prediction method in predicting the anticipated workloads in similar distributed environments will be one of our future research directions.

Competing Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants nos. 61502209 and 61502207 and the Natural Science Foundation of Jiangsu Province under Grant no. BK20130528.

References

[1] H. Al-Aqrabi, L. Liu, R. Hill, and N. Antonopoulos, "Cloud BI: future of business intelligence in the cloud," *Journal of Computer and System Sciences*, vol. 81, no. 1, pp. 85–96, 2015.

[2] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Performance evaluation of a green scheduling algorithm for energy savings in cloud computing," in *Proceedings of the IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW '10)*, pp. 1–8, Atlanta, Ga, USA, April 2010.

[3] L. Ceuppens, A. Sardella, and D. Kharitonov, "Power saving strategies and technologies in network equipment opportunities and challenges, risk and rewards," in *Proceedings of the International Symposium on Applications and the Internet (SAINT '08)*, pp. 381–384, August 2008.

[4] J. Li, B. Li, T. Wo et al., "CyberGuarder: a virtualization security assurance architecture for green cloud computing," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 379–390, 2012.

[5] J. Panneerselvam, L. Liu, N. Antonopoulos, and Y. Bo, "Workload analysis for the scope of user demand prediction model evaluations in cloud environments," in *Proceedings of the 7th IEEE/ACM International Conference on Utility and Cloud Computing (UCC '14)*, pp. 883–889, December 2014.

[6] Z. Wan, "Sub-millisecond level latency sensitive cloud computing infrastructure," in *Proceedings of the 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT '10)*, pp. 1194–1197, Moscow, Russia, October 2010.

[7] H. Zhang, G. Jiang, K. Yoshihira, H. Chen, and A. Saxena, "Intelligent workload factoring for a hybrid cloud computing model," in *Proceedings of the Congress on Services—I (SERVICES '09)*, pp. 701–708, 2009.

[8] C. Glasner and J. Volkert, "Adaps—a three-phase adaptive prediction system for the run-time of jobs based on user behaviour," *Journal of Computer and System Sciences*, vol. 77, no. 2, pp. 244–261, 2011.

[9] C. A. L. Fehling, Frank, Retter et al., "CloudComputingPatterns2014," 2014.

[10] J. Wang, G. Jia, A. Li, G. Han, and L. Shu, "Behavior aware data placement for improving cache line level locality in cloud computing," *Journal of Internet Technology*, vol. 16, no. 4, pp. 705–716, 2015.

[11] G. Han, W. Que, G. Jia, and L. Shu, "An efficient virtual machine consolidation scheme for multimedia cloud computing," *Sensors*, vol. 16, no. 2, article 246, 2016.

[12] S. Mahambre, P. Kulkarni, U. Bellur, G. Chafle, and D. Deshpande, "Workload characterization for capacity planning and performance management in IaaS cloud," in *Proceedings of the 1st IEEE International Conference on Cloud Computing for Emerging Markets (CCEM '12)*, pp. 1–7, Bangalore, India, October 2012.

[13] A. Khan, X. Yan, S. Tao, and N. Anerousis, "Workload characterization and prediction in the cloud: a multiple time series approach," in *Proceedings of the IEEE Network Operations and Management Symposium (NOMS '12)*, pp. 1287–1294, IEEE, Maui, Hawaii, USA, April 2012.

[14] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: insights from Google compute clusters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[15] P. A. Dinda and D. R. O'Hallaron, "Host load prediction using linear models," *Cluster Computing*, vol. 3, no. 4, pp. 265–280, 2000.

[16] S. Di, D. Kondo, and W. Cirne, "Google hostload prediction based on Bayesian model with optimized feature combination," *Journal of Parallel and Distributed Computing*, vol. 74, no. 1, pp. 1820–1832, 2014.

[17] I. S. Moreno, P. Garraghan, P. Townend, and J. Xu, "An approach for characterizing workloads in google cloud to derive realistic resource utilization models," in *Proceedings of the IEEE 7th International Symposium on Service-Oriented System Engineering (SOSE '13)*, pp. 49–60, IEEE, Redwood City, Calif, USA, March 2013.

[18] N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in

Proceedings of the IEEE 4th International Conference on Cloud Computing (CLOUD '11), pp. 500–507, Washington, DC, USA, July 2011.

[19] Z. Wan, "Cloud Computing infrastructure for latency sensitive applications," in *Proceedings of the IEEE 12th International Conference on Communication Technology (ICCT '10)*, pp. 1399–1402, November 2010.

[20] M. S. Bali and S. Khurana, "Effect of latency on network and end user domains in cloud computing," in *Proceedings of the 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE '13)*, pp. 777–782, Chennai, India, December 2013.

[21] Z. Wan, P. Wang, J. Liu, and W. Tang, "Power-aware cloud computing infrastructure for latency-sensitive internet-of-things services," in *Proceedings of the UKSim 15th International Conference on Computer Modelling and Simulation (UKSim '13)*, pp. 617–621, April 2013.

[22] C. Reiss, J. Wilkes, and J. L. Hellerstein, "Google cluster-usage traces: format + schema," Tech. Rep., Google Inc., Mountain View, Calif, USA, 2011.

[23] Google, "Google Cluster Data V1," 2011, https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md.

[24] J. Panneerselvam, L. Liu, N. Antonopoulos, and M. Trovati, "Latency-aware empirical analysis of the workloads for reducing excess energy consumptions at cloud datacentres," in *Proceedings of the IEEE 11th Symposium on Service-Oriented System Engineering (SOSE '16)*, pp. 62–70, Oxford, UK, March 2016.

[25] Z. Uykan, C. Güzeliş, and H. N. Koivo, "Analysis of input-output clustering for determining centers of RBFN," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, pp. 851–858, 2000.

[26] X. Sun, Z. Yang, and Z. Wang, "The application of BP neutral network optimized by genetic algorithm in transportation data fusion," in *Proceedings of the IEEE 2nd International Conference on Advanced Computer Control (ICACC '10)*, pp. 560–563, Shenyang, China, March 2010.

[27] W. C. Wang, *BP Neural Network and Application in Automobile Engineering*, BeiJing Institute of Technology University, 1998.

[28] Z. Li, Q. Lei, X. Kouying, and Z. Xinyan, "A novel BP neural network model for traffic prediction of next generation network," in *Proceedings of the 5th International Conference on Natural Computation (ICNC '09)*, pp. 32–38, Tianjin, China, August 2009.

[29] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*, PWS Publishing, Boston, Mass, USA, 1996.