A Review and Taxonomy on Data Driven Regression Models for Estimating Future Cloud Workloads

Dinesh Tulasidas Bharote[1], Prof. Pallavi Bagde[2]
Department of CSE,SDBCE, Indore[1,2]

***Abstract*—Cloud Computing has long become asought after fields in computer science. Several applications which need high computational complexity but cannot be performed on conventional hardware prefer to leverage cloud based platforms. Hence with increasing traffic and load on cloud servers or cloud based platforms, there seems to be a natural need for cloud workload prediction so as to estimate and manage cloud based resources. Since cloud data is large and complex at the same time, hence it is necessary to use artificial intelligence based techniques for the estimation of cloud workload so as to improve upon the accuracy of conventional techniques. This paper presents a review on the contemporary techniques for cloud workload prediction. The performance evaluation parameters have also been discussed.. It is expected that the paper would present with a headway for further research in cloud workload prediction.**

Keywords*—Cloud Workload Prediction, Aftificial Intelligence, Machine Learning, Artificial Neural Network (ANN), Mean Absolute Percentage error, Mean Square Error.

I. INTRODUCTION

Cloud Computing has revolutionized computational technology with cloud based platforms catering to the needs of systems unable to run complex processes on available hardware. The basic services provided by cloud computing are:

1) PAAS: Platform as Service
2) IAAS: Infrastructure as Service
3) SAAS: Software as service

With more sophisticated applications, it has become mandatory for tech giants to resort to cloud based services.

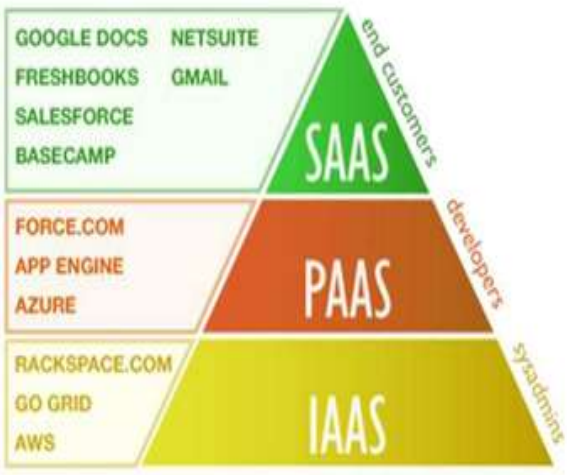

Fig.1 Cloud Services

With increasing number of users as well as large data sized, cloud workload has also seen a surge. Hence it is necessary to forecast cloud workload since several users try to access cloud services. However, the data being large and complex needs the aid of Artificial Intelligence for the prediction for the prediction purpose[4]. Cloud workload forecasting is typically challenging due to the number of users and the enormity of the data.

II. ARTIFICIAL NEURAL NETWORKS

Artificial Intelligence and Machine Learning (AI &ML) are preferred techniques for analyzing large and complex data. Generally, artificial neural networks (ANN) are used for the implementation of artificial intelligence practically. The architecture of artificial intelligence can be practically implemented by designing artificial neural networks. The biological-mathematical counterpart of artificial neural networks has been shown below.

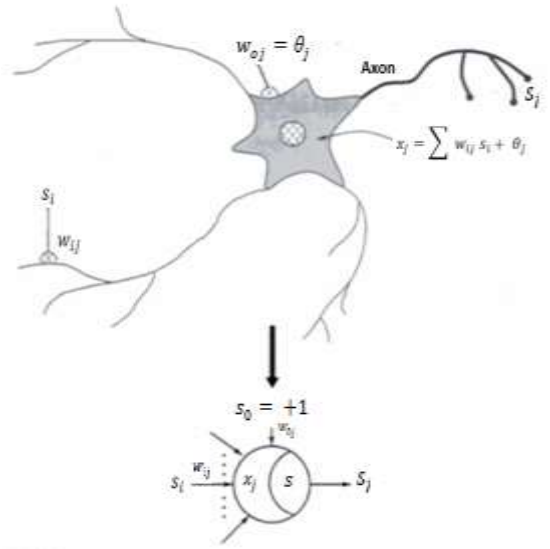

Fig.1 Biological-Mathematical Counterpart of ANN

The mathematical conversion of the ANN can be done by analyzing the biological structure of ANN. In the above example, the enunciated properties of the ANN that have been emphasized upon are:

1)Strength to process information in parallel way.
2) Learning and adapting weights
3)Searching for patterned sets in complex models of data.

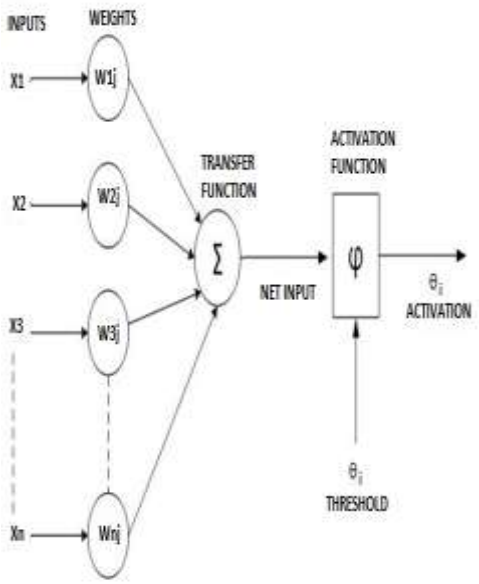

Fig.2 Mathematical Modeling of ANN

To see how the ANN really works, a mathematical model has been devised here, to indicate the functions mathematically.[7]. Here it is to be noted that the inputs of information parallel goes on into the input layer as specified whereas the end result analysis is marked from the output layer.

The feature of parallel acceptance and processing of data by the neural network serves a vital role. This ensures efficient and quicker mode of operation by the neural network. Also adding to it, the power to learn and adapt flexibly by the neural network aids in processing of data at a faster speed. **[2]**These great features and attributes make the ANN self dependent without requiring much intervention from humans. The output of the neural networks can be given by:

$$Y = \sum_{i=1}^{n} X_i . W_i + \theta_i \quad (1)$$

Here,

Y represents output

X represents inputs

W represents weights

Θ represents Bias

Training of ANN is of major importance before it can be used to predict the outcome of the data inputs. Neural Networks can be used for a variety of different purposes such as pattern recognition in large and complex data pattern sets wherein the computation of parameters would be extremely daunting for conventional statistical techniques. The weights or the equivalents of experiences are evaluated and updated based on the data patterns which are fed to the neural networks for training. Thus using artificial neural networks for the prediction or forecasting of cloud workload is an efficient proposition.

III. PREVIOUS WORK

This section highlights the prominent work in the domain.

Yazdanian et al. proposed a hybrid E2LG algorithm, which decomposes the cloud workload time-series into its constituent components in different frequency bands using empirical mode decomposition method which reduces the complexity and nonlinearity of prediction model in each frequency band. Also, a new state-of-the-art ensemble GAN/LSTM deep learning architecture is proposed to predict each sub band workload time-series individually, based on its degree of complexity and volatility. The ensemble GAN/LSTM architecture, which employs stacked LSTM blocks as its generator and 1D ConvNets as discriminator, can exploit the long-term nonlinear dependencies of cloud workload time-series effectively specially in high-frequency, noise-like components

Gao et al. showed that meeting QoS with cost-effective resource is a challenging problem for CSPs because the workloads of Virtual Machines (VMs) experience variation over time. It is highly necessary to provide an

accurate VMs workload prediction method for resource provisioning to efficiently manage cloud resources. In this paper, authors first compare the performance of representative state-of-the-art workload prediction methods. We suggest a method to conduct the prediction a certain time before the predicted time point in order to allow sufficient time for task scheduling based on predicted workload. To further improve the prediction accuracy, authors introduce a clustering based workload prediction method, which first clusters all the tasks into several categories and then trains a prediction model for each category respectively. The trace-driven experiments based on Google cluster trace demonstrates that our clustering based workload prediction methods outperform other comparison methods and improve the prediction accuracy to around 90% both in CPU and memory.

Chen et al. proposed a deep Learning based Prediction Algorithm for cloud Workloads (L-PAW). First, a top-sparse auto-encoder (TSA) is designed to effectively extract the essential representations of workloads from the original high-dimensional workload data. Next, authors integrate TSA and gated recurrent unit (GRU) block into RNN to achieve the adaptive and accurate prediction for highly-variable workloads. Using real-world workload traces from Google and Alibaba cloud data centers and the DUX-based cluster, extensive experiments are conducted to demonstrate the effectiveness and adaptability of the L-PAW for different types of workloads with various prediction lengths. Moreover, the performance results show that the L-PAW achieves superior prediction accuracy compared to the classic RNN-based and other workload prediction methods for high-dimensional and highly-variable real-world cloud workloads.

Wang et al. provided Adaptive Dispatching of Tasks in Cloud. Cloud computing domain has been witnessing a large traffic and users dependent on it. With most of the work being shifted to the internet platform, the cloud services have become dominant in all aspects of business and technology. In this work, the authors proposed a novel study of cloud tasks dispatching. There are allocation schemed and algorithms that have been used as a part of the model. The response time that is computed has been reduced considerably in this work. Various hosts have been deployed for the proper client and server interaction. The time delays were greatly lessened and it proved to be a really useful methodology.

Duggan et al. presented Research on Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks. This study aims to predict the CPU consumption of host machines by using recurrent neural networks. The process involved utilizing the recurrent neural networks that could accurately predict the time series data and also collect the information with flexibility. With respect to the traditional approaches and methods, this method was successful in accurate forecasting and gave better outcomes.

Liu et al. propose A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning. It stood for a novel hierarchical framework that could address and solve all the possible power and resource allocation problems in the cloud based platforms. The proposed system took into account the virtual machines servers and various resources. The rising use of the reinforced deep learning solutions also helped in restructuring the entire concept and model. The workload prediction could be used for several other formats and henceforth is great way to rebuild the systems. The outcomes were overlaid and all the required resources were outsourced to the proposed model.

Zuo et al. proposed A Multiqueue Interlacing Peak Scheduling Method Based on Tasks Classification in Cloud Computing. It was mainly a scheduling scheme that was further improved. The resource allocation and tasks classification was carried out on the basis of the type of memory and the CPU consumption. The infrastructure within the workloads may vary. Put together, they give rise to a complete cloud solution. The CPU specific tasks were classified differently and the peak scheduling was used for it. The interlacing was found to be useful for all the separate parts of the processing model. It could be used well with their other counterparts. Overall it was a very robust mechanism that provided great accuracy in classification and added efficacy to the complete system.

Hu et al. propose Three Models to Predict the Workload Based on Analysing Monitoring Data. The dataset for the cloud workload is a very important part of gauging the entire system design. The authors proposed three models for forecasting the cloud workload. And the help was taken from the dataset for the workload. By monitoring the data and information flow, it is easy to predict the workload extent and its quantity. This helps in building elasticity and also enhances the scalability of the system. The workload plays a crucial role and it must be flexible enough so that different programs can use it according to its changing requirements. The concept of Cloud Workload prediction can help in determining the cloud storage and it's planning for different applications with ease.

Xue et al. put forth PRACTISE, a neural network based framework that could predict the future cloud workloads, peak loads etc. The cloud workload prediction has been a very active area of research and the authors primarily focused on forecasting the peak loads and their timings etc. As due to overflow of data and resources, the cloud serves hold the probability to crash and go off. So, forecasting helps in giving optimization solutions to the problems faced. This approach worked well with the methods and offered improved accuracy and elasticity.

Abdelwahab et al. Enabling Smart Cloud Services through Remote Sensing: An Internet of Everything Enabler survey. The concept of remote sensing has been utilized in this research work. The use of the cloud services by the IoT and remote sensing techniques has been the subject of the study. It is a very good concept to use both the technologies together. Already the emergence of the IoT has been coupled with the cloud based services and their amalgamation has been quite a success. This survey points out the aspects of the remote sensing for cloud services and other allied areas where it can be applied. Cloud based platforms provide several applications such as web services, security services, big data and machine learning services.

This section presents a comprehensive review of the different previous approaches in terms of the various different models designed and the associated parameters. A tabulation of the contemporary work along with the approaches has been given below:

Authors	Approach
Yazdanian et al.	A deep learning based Long Short Term Memory Based approach for cloud workload forecasting.
Gao et al.	A machine Learning based trace-driven experiments based on Google cluster trace demonstrates that our clustering based workload prediction methods outperform other comparison methods and improve the prediction accuracy to around 90% both in CPU and memory.
Jitendra Kumar et al.	The approach used for cloud workload prediction used in this paper is neural networks in conjugation with differential evolution approach
Lan Wang et al.	This approach focusses on the dispatch of tasks and load on the cloud server as an optimization problem
Martin Duggan et al.	The proposed approach presents a forecast of CPU utilization of Cloud servers using recurrent neural network learning.
Ning Liu et al.	This approach uses a reinforcement learning in neural networks for resource management in cloud servers.
Liyun Zuo et al.	The approach uses a multi queue based interlacing approach for classification of cloud computing services.
Yazhou Hu et al.	The approach focusses on cloud workload estimation and the performance metrics is accuracy. The approach tests the model for different data sets.
Ji Xue et al.	The proposed approach predicts the number of data centers in cloud based servers based on the analysis of previous data.
Mehmet Demirci	The paper presents a comprehensive survey on machine learning based approaches for cloud resource management.
Sherif Abdelwahab et al.	The approach proposed an internet of everything based approach using cloud based services to connect devices over internet
Chin-Feng Lai et al.	The approach uses a collaborative computing based approach using cloud based servers for Wireless Body Sensor Network applications.

Table.1 Comparative Analysis of Previous Work

IV. PERFORMANCE METRICS

Since the purpose of the proposed work is time series prediction, hence it is necessary to compute the required performance metrics. Since there is a chance of positive and negative errors to cancel out, hence it is necessary to compute the Mean Absolute Percentage Error (MAPE) given by:

$$MAPE = \frac{100}{M}\sum_{t=1}^{N}\frac{E-E_t|}{E_t} \quad (2)$$

Here,
N is the total number of samples
E is the actual value
$\mathbf{E_t}$ is the predicated value

The mean square error is also evaluated often to stop training, which is given mathematically by:

$$MSE = \frac{1}{N}e_i^2 \quad (3)$$

Here,
E is the error
N is the number of samples

It is always envisaged to attain low error values and high values of accuracy for cloud workload prediction.

CONCLUSION

The present work renders insight into the basic methodologies working as empirical models for cloud load forecasting as a time series prediction Cloud services have brought a radical shift in the computing domain with loads of benefits for users who want quality services and functionalities in machines. The Cloud services generally operate based on big server based machines that can provide services according to the user requirements and requests. The load of work and services on the cloud can vary depending upon the demands and requests. So Prediction of the work load can be of major use for the optimization of the cloud efficacy.

REFERENCES

[1] P Yazdanian, S Sharifian, E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction", Journal of Supercomputing, Springer 2022, vol. 77, pp.11052–11082.

[2] J. Gao, H. Wang and H. Shen, "Machine Learning Based Workload Prediction in Cloud Computing," 2020 29th International Conference on Computer Communications and Networks (ICCCN), 2020, pp. 1-9

[3] Z. Chen, J. Hu, G. Min, A. Y. Zomaya and T. El-Ghazawi, "Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning," in IEEE Transactions on Parallel and Distributed Systems, 2020, vol. 31, no. 4, pp. 923-934.

[4] L. Wang and E. Gelenbe, "Adaptive Dispatching of Tasks in the Cloud," in IEEE Transactions on Cloud Computing, vol. 6, no. 1, pp. 33-45, 1 Jan.-March 2018

[5] Martin Duggan, Karl Mason, Jim Duggan, Enda Howley, Enda Barrett, "Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks", 2017 IEEE.

[6] Ning Liu, Zhe Li, Jielong Xu, Zhiyuan Xu, Sheng Lin, Qinru Qiu, Jian Tang, Yanzhi Wang, "A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning", 2017 IEEE.

[7] Liyun Zuo, Shoubin Dong, Lei Shu, Senior Member, IEEE, Chunsheng Zhu, Student Member, IEEE, and Guangjie Han, Member, IEEE, "A Multiqueue Interlacing Peak Scheduling Method Based on Tasks' Classification in Cloud Computing", 2016 IEEE.

[8] Yazhou Hu, Bo Deng, Fuyang Peng and Dongxia Wang, "Workload Prediction for Cloud Computing Elasticity Mechanism", 2016 IEEE.

[9] Ji Xue, Feng Yan, Robert Birke, Lydia Y. Chen, Thomas Scherer, and Evgenia Smirni, "PRACTISE: Robust Prediction of Data Center Time Series", 2015 IEEE.

[10] Mehmet Demirci, "A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments", 2015 IEEE.

[11] Sherif Abdelwahab, Member, IEEE, Bechir Hamdaoui, Senior Member, IEEE, Mohsen Guizani, Fellow, IEEE, and Ammar Rayes, "Enabling Smart Cloud Services Through Remote Sensing: An Internet of Everything Enabler", 2014 IEEE.

[12] Chin-Feng Lai, Member, IEEE, Min Chen, Senior Member, IEEE, Jeng-Shyang Pan, Chan-Hyun Youn, Member, IEEE, and Han-Chieh Chao, Senior Member, IEEE, "A Collaborative Computing Framework of Cloud Network and WBSN Applied to Fall Detection and 3-D Motion Reconstruction", 2014 IEEE.

[13] Ian Davis, Hadi Hemmati, Ric Holt, Mike Godfrey, Douglas Neuse, Serge Mankovskii, "Storm Prediction in a Cloud", 2013 IEEE.

[14] Abul Bashar, "Autonomic Scaling of Cloud Computing Resources using BN-based Prediction Models", 2013 IEEE.

[15] Sadeka Islam , Jacky Keunga, Kevin Lee, Anna Liu, "Autonomic Scaling of Cloud Computing Resources using BN-based Prediction Models", 2012 ELSEVIER.

[16] Erol Gelenbe, Ricardo Lent and Markos Douratsos, "Choosing a Local or Remote Cloud", 2012 IEEE.

[17] Mohammad Moein Taheri and Kamran Zamanifar, "2-Phase Optimization Method for Energy Aware Scheduling of Virtual Machines in Cloud Data Centers", 2011 IEEE.

[18] Saurabh Kumar Garg, Srinivasa K. Gopalaiyengar, and Rajkumar Buyya, "SLA-Based Resource Provisioning for Heterogeneous Workloads in a Virtualized Cloud Datacenter", 2011 IEEE.

[19] Md. Toukir Imamt, Sheikh Faisal Miskhatt, Rashedur M Rahmant, M. Ashraful Amin, "Neural Network and Regression Based Processor Load Prediction for Efficient Scaling of Grid and Cloud Resources", 2011 IEEE.