A Review on Soft Computing Approaches for Workload Forecasting for Cloud Services

[1]Richa Pandey, [2]Prof.Narendra Kumar,

[1]Research Scholar, [2]Assistant Professor,
[1]Department of Computer Science, [2] Department of Computer Science ,
[1]Institute of Engineering, SAGE University, Indore, India.

Abstract : Soft Computing based approaches are being used today in several domains for data analytics, one of which is cloud computing too. Cloud Computing has long become a sought 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 soft computing based techniques for cloud workload forecasting. 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.

***IndexTerms*–Soft Computing, Cloud Computing, Cloud Workload, Workload Estimation, Mean Absolute Percentage Error (MAPE)**

I. INTRODUCTION

Soft Computing based approaches are being widely used for big data analytics where human intervention and conventional techniques do not suffice. Typically, soft computing based approaches try to mimic the human thought process which in in contrast to the conventional pre-programmed behavior or computational platforms. The basic soft computing approaches used for cloud workload predictions, as a time series prediction problem are:

1) Neural Networks
2) Fuzzy Logic
3) Neuro Fuzzy Expert Systems

The basics of these approaches and their need in cloud workload prediction are discussed in the subsequent section:

1.1 Neural Networks

Neural networks try to mimic the parallel data processing and self-adapting performance of the human brain. 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.

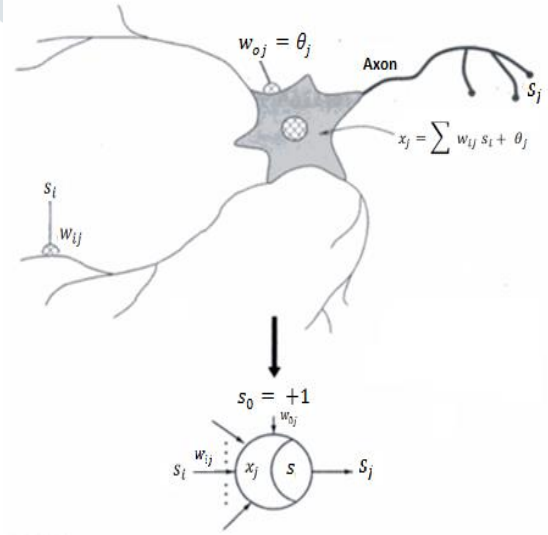

Figure.1 Biologic and Mathematical Equivalents of Neural Networks

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.

1.2 Fuzzy Logic

The need for fuzzy logic or fuzzy systems arise when there is a sense of uncertainty on the determination of a category of data belonging to a larger dataset. The analogy of clear and unclear boundaries illustrate the concept.

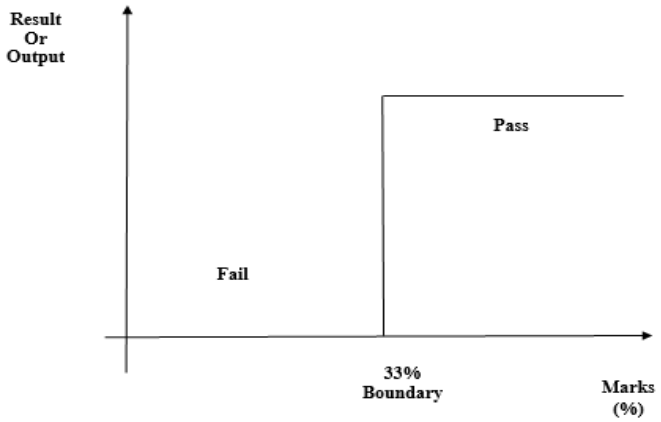

Figure.2 Illustration of a clear boundary

The figure above depicts the clear boundary or marker to differentiate data sets. Here, the boundary can be mathematically defined as:

$$Y(x) = L1; x \geq T \quad (1.1)$$

And

$$Y(x) = L2; x < T \quad (1.2)$$

Here,
Y is the output
T is the boundary or threshold
L1 is level one
L2 is level 2.
However, this sort of a clear boundary may be non-existent in several cases such as the one illustrated below:

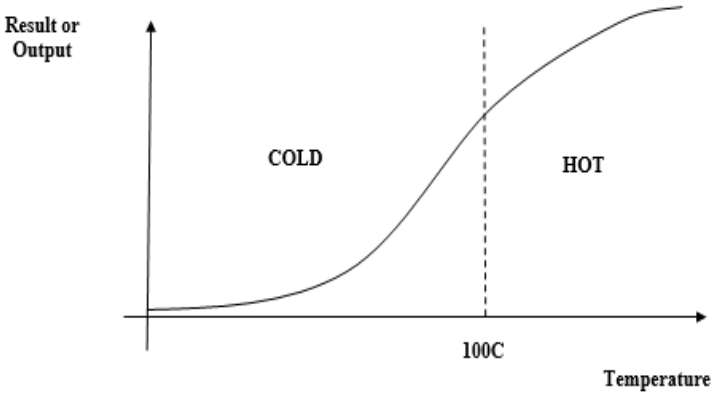

Figure.3 Concept of Fuzziness or Unclear boundary

The above figure clearly depicts that there is no clear boundary between the dataset but a fuzzy boundary.

1.3 Neuro Fuzzy Expert Systems

The neuro-fuzzy expert systems is a combination of:

- Neural Networks
- Fuzzy Logic

In this case, the membership functions of the fuzzy inference machine are governed by a neural network. The following figure depicts the neuro fuzzy expert system architecture.

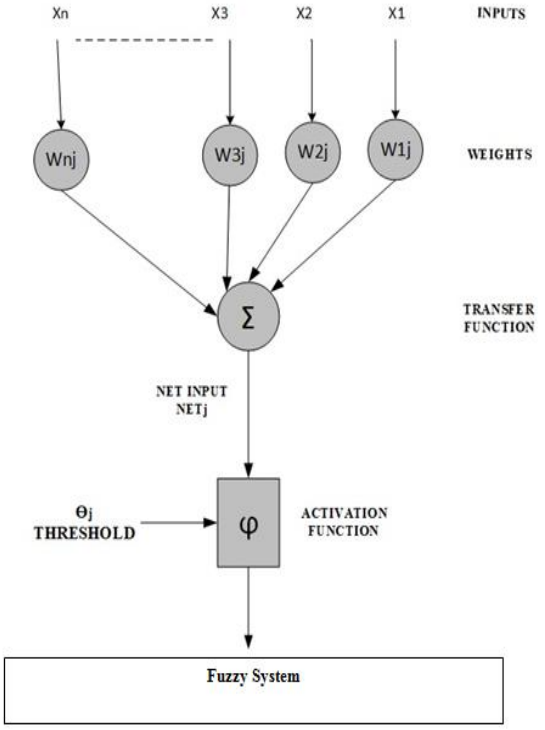

Figure.4 Architecture of a neuro fuzzy expert system

The output of the neural network is applied to the fuzzy block which finally renders the output.

II. Cloud Services

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.

Figure.1 Fundamental Cloud based Services

With increasing number of users as well as large data sized, cloud workload has also seen a surge [1]. 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. Cloud workload forecasting is typically challenging due to the number of users and the enormity of the data from the hidden layer.

III. Previous Work

This section highlights the prominent work in the domain.

In Elsevier 2018, Jitendra Kumar et al. [1] proposed the estimate of workload on cloud based platforms using ANN. The approach used the concept of crossover with the mutation that was to attain better results compared to conventional mutation. The data sets used were that of NASA and Saskatchewan servers' HTTP traces for different prediction intervals. The major gap that could be seen in this approach was the fact that there was no error feedback path leading to better error reduction. Moreover, there

wasn't a comprehensive analysis on the time complexity in terms of the number of iterations for the model to train so as to render an idea about the actual time needed for the real time critical applications.

In IEEE 2018, Lan Wang et al. [2] proposed a technique that used online QoS as an information metric for the optimization of load balancing and task scheduling. The major contribution of the work was the task scheduler model for the given data. The major gap or limitation of the approach was the fact that the task scheduling designed was not designed to incorporate the predictive model's inputs. The system designed used a simple round-robin scheduler which is again an empirical model for scheduling with no predictive measures for enhanced performance of the scheduling algorithm.

In IEEE 2017, Martin Duggan et al. [3] designed a model that could predict the CPU consumption for cloud based platforms based on a recurrent neural network (RNN) architecture. The major gap in the Recurrent Neural Network can be thought of as the lack of processing speed of large chunks of data in real time critical applications. The CPU scheduling based on CPU utilization does not give nay idea about the actual load that the cloud based servers may need to face. Moreover the RNN approach does not utilize the steepest descent approach for the gradient computation making the time complexity more.

In IEEE 2017, Ning Liu et al. [4] proposed a model for managing the resource allocation of the system. The approach also tries to reduce the power consumption of the system. The proposed system's global tier uses the deep reinforcement learning (DRL) technique while the local tier of uses the LSTM based workload predictor model. The major limitation of using a two tier approach for resource management and workload prediction is the fact that the a two tier approach often called the ada boost or the bagging concept in neural networks increase the computational complexity of the system to a large extent and this makes the system to be infeasible in real time applications. Again, this would mean that one training and testing algorithm would not suffice for the diversity of the cloud workloads of different sizes and varieties. The more effective approach could however be using a single neural network with multiple hidden layers for enhanced data processing with low execution time which can be clearly gauged in terms of the number of epochs.

In IEEE 2016, Liyun Zuo et al. [5] proposed a system for scheduling of tasks on cloud based platforms depending on the system parameters which in case of the virtual machines are response time, load balancing deadline violation and resource utilization. The major limitation of a CPU-intensive and IO-intensive approach is the fact that there may occurrence of a sudden migration of the VM form CPU intensive teaks to IO intensive teaks which can happen back and forth. The approach doesn't use any sophisticated interrupt or interrupt timer to merge the CPU and IO intensive scheduling.

In IEEE 2016, Yazhou Hu et al. [6] proposed an approach using time series prediction and the Kalman Filter based approach with a main aim of decreasing automatic scaling delay (ASD). The major gap in using a Kalman Filter based approach for time series prediction is the fact that the approach doesn't use a feedback mechanism for errors which can help in weight updation. The approach lacks the decision about the steepest descent and hence doesn't allow steepm reduction in the prediction error.

In IEEE 2015, Ji Xue et al. [7] proposed a neural network based approach that could predict future loads, peak loads, and their timing. The data set used was that of the IBM data centre. The proposed approach when compared to the standard ARIMA tool was found to work better. The major challenge with using simple benchmark neural network models is the fact that such an approach doesn't cater to the need to the design of different neural models for different sets of data. Moreover, no data pre-processing mechanism is used for pre-processing the raw data prior to feeding to the benchmark models.

In IEEE 2015, Mehmet Demirci [8] present a survey on various machine learning based approaches for prediction of cloud workload, The limitations found in the survey is the fact that effective learning of neural networks is necessary to learn from past patterns and hence be able to predict future workloads. Due to the enormity and the irregular or random nature of cloud workload data, machine learning techniques are necessary with customized adaptive learning for cloud VMs.

In IEEE 2014, Sherif Abdelwahab et al [9] proposed a system for smart cloud service using remote sensing for internet of everything or internet of things (IoT) based applications. The predictor model is based on sensing the previous data for IoT based cloud platforms. The major gap in the IoT based cloud model is the fact that the there is no provision to analyze the data on real time gadgets. This makes several real time applications infeasible.

In IEEE 2014, Chin-Feng Lai et al [10] proposed an interesting framework using cloud computing and wireless body area network. The data sensed by the WBAN is dent to the cloud platform for analysis and further controlling. The major gap in the approach of Wireless Networks intertwined with Cloud Computing platforms is the fact that there is no provision for rendering the security to such systems. The authentication of user data can be a serious challenge with adversaries manipulating data or even performing man in the middle attacks on the system

The limitations found in the survey is the fact that cloud workload forecasting is majorly challenging due to the largeness and magnanimity of the cloud data sets. Due to the enormity and the irregular or random nature of cloud workload data, machine learning techniques are necessary with customized adaptive learning for cloud VMs.. The major gap in the Recurrent Neural Network can be thought of as the lack of processing speed of large chunks of data in real time critical applications. The CPU scheduling based on CPU utilization does not give nay idea about the actual load that the cloud based servers may need to face. Moreover the RNN approach does not utilize the steepest descent approach for the gradient computation making the time complexity more.

The major limitation of using a two tier approach for resource management and workload prediction is the fact that the a two tier approach often called the ada boost or the bagging concept in neural networks increase the computational complexity of the system to a large extent and this makes the system to be infeasible in real time applications. Again, this would mean that one training and testing algorithm would not suffice for the diversity of the cloud workloads of different sizes and varieties. The significant challenges in the design of a particular algorithm lies in the fact that the approach has to meet the constraints of space and time complexity both, This indirectly means that the approach has to have low execution time and stability in reduction of errors as the weights change wit each iteration. The number of iterations or epochs needed in the training would be an indicator which would be accountable for inferring the time complexity.

IV. EVALUATION PARMAETRES

The 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:

$$MAPE = \frac{100}{M}\sum_{t=1}^{N}\frac{E-E_t|}{E_t} \quad (4.1)$$

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:

$$MSE = \frac{1}{N}e_i^2 \quad (4.2)$$

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.

IV. 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] Jitendra Kumar , Ashutosh Kumar Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution", 2018 Elsevier.
[2] Lan Wang and Erol Gelenbe, "Adaptive Dispatching of Tasks in the Cloud", 2018 IEEE.
[3] Martin Duggan, Karl Mason, Jim Duggan, Enda Howley, Enda Barrett, "Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks", 2017 IEEE.
[4] 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.
[5] Liyun Zuo, Shoubin Dong, Lei Shu, , IEEE, Chunsheng Zhu, Guangjie Han,"A Multiqueue Interlacing Peak Scheduling Method Based on Tasks' Classification in Cloud Computing", 2016 IEEE.
[6] Yazhou Hu, Bo Deng, Fuyang Peng and Dongxia Wang, "Workload Prediction for Cloud Computing Elasticity Mechanism", 2016 IEEE.
[7] Ji Xue, Feng Yan, Robert Birke, Lydia Y. Chen, Thomas Scherer, and Evgenia Smirni, "PRACTISE: Robust Prediction of Data Center Time Series", 2015 IEEE.
[8] Mehmet Demirci, "A Survey of Machine Learning Applications for Energy-Efficient Resource Management in Cloud Computing Environments", 2015 IEEE.
[9] Sherif Abdelwahab, Bechir Hamdaoui, Mohsen Guizani, Ammar Rayes, "Enabling Smart Cloud Services Through Remote Sensing: An Internet of Everything Enabler", 2014 IEEE.
[10] Chin-Feng Lai, Min Chen, Jeng-Shyang Pan, Chan-Hyun Youn, Han-Chieh Chao, "A Collaborative Computing Framework of Cloud Network and WBSN Applied to Fall Detection and 3-D Motion Reconstruction", 2014 IEEE.
[11] Ian Davis, Hadi Hemmati, Ric Holt, Mike Godfrey, Douglas Neuse, Serge Mankovskii, "Storm Prediction in a Cloud", 2013 IEEE.
[12] Abul Bashar, "Autonomic Scaling of Cloud Computing Resources using BN-based Prediction Models", 2013 IEEE.
[13] Sadeka Islam , Jacky Keunga, Kevin Lee, Anna Liu, "Autonomic Scaling of Cloud Computing Resources using BN-based Prediction Models", 2012 ELSEVIER.
[14] Erol Gelenbe, Ricardo Lent and Markos Douratsos, "Choosing a Local or Remote Cloud", 2012 IEEE.
[15] Mohammad Moein Taheri and Kamran Zamanifar, "2-Phase Optimization Method for Energy Aware Scheduling of Virtual Machines in Cloud Data Centers", 2011 IEEE.