www.ijaceeonline.com ISSN: 2456 - 3935

International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

A Flexible Approach for Predicting CPU Utilization for Cloud Workloads

Padma D. Adane
Associate Professor, Department of Information Technology,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India
Email: adanep@rknec.edu

Omprakash G. Kakde
Professor and Head, Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology, Nagpur, India
Email: ogkakde@cse.vnit.ac.in

Abstract: *Workload Prediction is an important step performed by any proactive provisioning technique. The efficiency of such techniques rests highly on the prediction module that foresees the future workload requirements and alerts the system to be prepared for the resource needs thereby reducing the resource provisioning time at the time of actual need. Using a single model for load prediction may give good prediction accuracy for a particular application. However, this approach may not be suitable in cloud environments where the workload patterns change dynamically. The current trend in research is towards the use of hybrid methods that combine the performance of individual prediction methods. This paper presents a flexible approach which uses two machine learning algorithms, Linear Regression and Support Vector Machine. Linear Regression adapt well to the linear changes in the workload while Support Vector Machine can handle the non-linear service workload changes in a cloud environment with better accuracy. The experiment results show that the combined approach performs better than the individual prediction model.*

Keyword: *Linear Regression; Machine Learning; Prediction Model; Support Vector Machine; Workload Prediction*

1. INTRODUCTION

On demand provisioning of resources is an important feature of cloud computing [1]. Provisioning of resource is the first phase of resource management [2] in a cloud environment that deals with finding the resource requirements of application for its deployment in the cloud. It is basically a high level resource management task which should adapt dynamically to the sudden rise and dip in the user's requirements. Efficient resource provisioning to meet the requirement of quality of service is a growing challenge [3].

There are two approaches to provision resources in a cloud environment: Reactive and Proactive. Reactive provisioning approaches are simple to implement. However, the overall response time is high as the provisioning decisions are taken when the need actually arises. On the other hand, proactive provisioning mechanisms are complex but response time is improved as the decisions are taken before the actual need arises. However, these provisioning mechanisms rely on a prediction model to foresee the future workload demands and make available the required resources quickly. As stated by Adrian Cockcroft (Technology Fellow at Battery Ventures, Ex - Cloud Architect at Netflix) in his keynote address at the 7th IEEE/ACM conference on Utility and Cloud Computing, December 2014, "Predictive auto-scaling saves up to 70% of cloud costs." An efficient proactive provisioning is possible only if the technique employs a good workload predictor that can forecast the future workload requirements and timely alert the system to make available the required resources as per the need, thereby, minimizing the provisioning time.

The modeling approaches for the application workload prediction can be broadly divided into four groups [4]: Table driven, Control theory, Queuing theory and Machine Learning techniques. The common machine learning techniques used by researchers for prediction are K-Nearest Neighbor, Linear Regression, Bayesian Model, Markov Model, Random Forest, Neural Network and Support Vector Machine. The conventional approach for prediction is to use a single model for load prediction which may give good prediction accuracy for a particular application. How-

Cite this paper:

Padma D. Adane, Omprakash G. Kakde, "A Flexible Approach for Predicting CPU Utilization for Cloud Workloads", International Journal of Advances in Computer and Electronics Engineering, Vol. 3, No. 9, pp. 7-11, September 2018.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

ever, this approach may not be suitable in cloud environments where the workload patterns change dynamically. As per the evaluation results in [5, 6], a single prediction method fails to provide accurate results for service workload changes in a cloud environment. There is a growing trend among researchers to use hybrid prediction methods which combine the efficiency of individual prediction models. In this paper we present one such combined approach using Linear Regression and Support Vector Machines.

Linear Regression (LR) is most basic method used in statistics where all the attributes involved in the prediction are numeric [7]. The output to be predicted is expressed as a linear combination of other involved attributes with predetermined weights. Equation 1 given below represents the value of the output to be predicted [8]:

$$y = a0 + a1 *x1 + a2 * x2 + \ldots + an * xn \quad (1)$$

where y is the output to be predicted, x1 to xn are the attributes involved in the prediction and a1 to an are the weights associated with these attributes. These weights are calculated from the training data. For data logs with highly correlated attributes, LR performs with reduced accuracy [9]. Moreover, for long-term forecasting, statistical methods perform poorly [10].

Support Vector Machine (SVM) is a machine learning technique that can be used for both classification as well as regression problems. When used for regression, it is referred as Support Vector Regression (SVR). SVR tries to minimize the error by finding a line of best fit [9]. To achieve this, it considers data instances nearest to the minimum cost line. Such instances are known as Support Vectors. To accommodate curved lines or polygon regions, it scales the data into higher dimensions for predictions. This can be achieved by trying out different kernels. For linear kernel the output to be predicted is calculated as given by Equation 2:

$$Y = b0 + \text{sum}\,(ai * (x, xi)) \quad (2)$$

where x is the input vector, xi are the support vectors. The coefficients b0 and ai (for each input) must be calculated from the training data. As suggested in [11], Support Vector Machines accurately forecast nonlinear, non-stationary times series data. It out performs other nonlinear prediction techniques like Neural Network based Multilayer perceptron.

In this paper, we have first evaluated the performance of these machine learning techniques individually for predicting the CPU utilization on five server logs taken from the public cloud server Parallel Workloads Archive [12] with Mean Absolute Error as the metrics. We have then evaluated the performance of an algorithm using combined approach for the same metrics. The results show that this flexible algorithm performs better than the individual models.

2. RELATED WORK

Currently, the direction of research is towards hybrid prediction methods that integrate the effectiveness of individual prediction model and provide better prediction results [4]. In [13], Neural Network and Linear Regression have been used for an improvised prediction based resource management and provisioning strategy. In order to provide accurate forecasting ahead of time, evaluation has been carried out with varying sliding window sizes. The data set used for evaluation is taken from TPC-W benchmark in the amazon EC2 cloud. Their results demonstrate the effectiveness of Neural Network model in forecasting the resource utilization. Authors in [14] have classified the service workload as fast time scale data and slow time scale data based on workload dynamics and used Linear Regression to predict slow time scale workloads and Support Vector Machine to predict fast time scale workloads. Evaluation has been carried out on real traces from in house developed application systems. The approach needs to be tested for different workloads in service cloud scenario.

In [15], the authors have used regression based simple prediction models to predict the resource requirements. These results were then fed to a fuzzy neural network with self adjusting learning rate and momentum weights as inputs to estimate the final resource demands. The approach has been evaluated for four different matrices and demonstrates the improvement in prediction performance. However, the use of two layered approach for prediction increases the prediction time.

In [10], authors have integrated statistical and machine learning methods for efficient workload prediction in cloud. They have extracted the features using a two phase pattern matching method and then used Random Forest for workload prediction and addressed the problem with classification as well as regression. The proposed model has been evaluated on the AuverGrid workload data series taken from the Grid Workloads Archive. The results obtained validate the prediction accuracy of the approach. For improving the resource management, [16] have combined the machine learning techniques with a single entity vision of the grid with an aim to ease fault tolerance and job scheduling in grid management.

In [17], authors have used an optimal decision recently strategy to present a self adaptive prediction algorithm that combines Linear Regression and BP neural network. The algorithm performs better with a higher accuracy on workload prediction compared to the individual methods. The algorithm has been further used to design a dynamic scheduling architecture to improve resource utilization. [18] has adopted a weighted linear combination strategy to combine the predictions of several methods. Starting with initial

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

TABLE I THE SERVER LOGS FROM THE PARALLEL WORKLOAD ARCHIVE

S. No.	Server Log	Total Number of instances taken for experiment	Description	Log Providers
1	SANDIA-ROSS	9521	CPlant Cluster with 48 cabinets of 32 nodes. This log contains three years of accounting records from the Sandia Ross cluster.	Jon Stearly
2	LANL-O2K	72200	Cluster of 16 Origin 2000 machines with 128 processors each. This log contains about 4 months of data derived from the accounting records produced by the LSF software running at Los Alamos National Lab.	Fabrizio Petrini
3	LANL-CM5	79425	1024-node Connection Machine CM-5 from Thinking Machines. This log contains two years of accounting records produced by the DJM software at Los Alamos National Lab.	Curt Canada
4	HPC2N	160177	120 –node Linux Cluster. This log contains three and a half years of accounting records from the High-Performance computing Centre North in Sweden.	Ake Sangren, Michael Jack
5	LPC-EGEE	99755	A cluster composed of 70 dual 3GHz Pentium-IV Xeons nodes running Linux. This log contains 10 months of records of Laboratory of Corpuscular Physics of University Blaise-Pascal, Clermont-Ferrand, France.	Emmanuel Medernach

equal weights, the approach updates the weights as per the prediction error of each method. The paper also introduces a novel asymmetric cost measure called Cloud Prediction Cost. The ensemble prediction model has been evaluated on the IBM smart Cloud Enterprise trace data. The presented result establishes the effectiveness of the model in improving the service quality.

3. EXPERIMENTAL SETUP AND RESULTS

We have performed the experiments in two steps. In the first step we have evaluated the individual performance of Linear Regression and Support Vector Machine in predicting the CPU utilization of different server logs. Next we have evaluated the performance of the same logs using the flexible algorithm. For these evaluations, we have used Python 3.6 and Intel Xeon Dell server with 3.2 GHz speed and 16 GB RAM. The server logs considered are in Standard Workload Format [12] with each instance/row consisting of 18 attributes.

Table I describes the logs that have been used for this work. Each data log has been normalized and prepossessed by sampling both horizontally and then vertically. Vertical sampling method generates inputs which come close to real-world usage of machine learning algorithms [19]. Each subset of attributes is evaluated with the target machine learning algorithm. The subset of attributes with optimal performance, low dimensionality and domain knowledge bias is chosen for prediction purposes. This pre-processing reduced the number of attributes considered for this evaluation to 5 comprising of wait time, run time, number of allocated processor, average CPU time used and used memory.

For horizontal sampling, cross-validation [20, 21] technique has been applied for estimating the accuracy of the prediction models. Cross-validation is one of the most common error estimation techniques where each observation in the sample dataset of size n is successively taken out and the remaining n−1 observations of the set are used to train the prediction model to estimate the projected resource usage [13]. The metrics used for evaluation is Mean Absolute Error (MAE) [22] which is defined as average of the absolute errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \hat{y}_i - y_i | \quad (3)$$

Equation 3 gives the formula for Mean Absolute Error where, y_i is the actual output, $\hat{y}_i$ is the predicted output and n is the number of observations in the dataset for which the prediction is made. Table II presents the MAE observations made for each of the server logs for the two machine learning algorithms. Figure 1 illustrates the graph for the obtained predictions for CPU utilization for the five server logs under consideration. For data with linear changes, Linear Regression performs better. While for dataset with non linear changes, prediction accuracy of Support Vector Machine is better. Thus, for dataset with different characteristics, the two methods have different prediction accuracy.

Table III gives the pseudo code of the flexible algorithm that combines the two methods. We have first normalized each data set so as to bring the value of each attribute in consideration in the range between 0 and 1. This method has been adopted as the distribution of the data in each log is not known. Then for each fold in consideration, we have calculated the

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

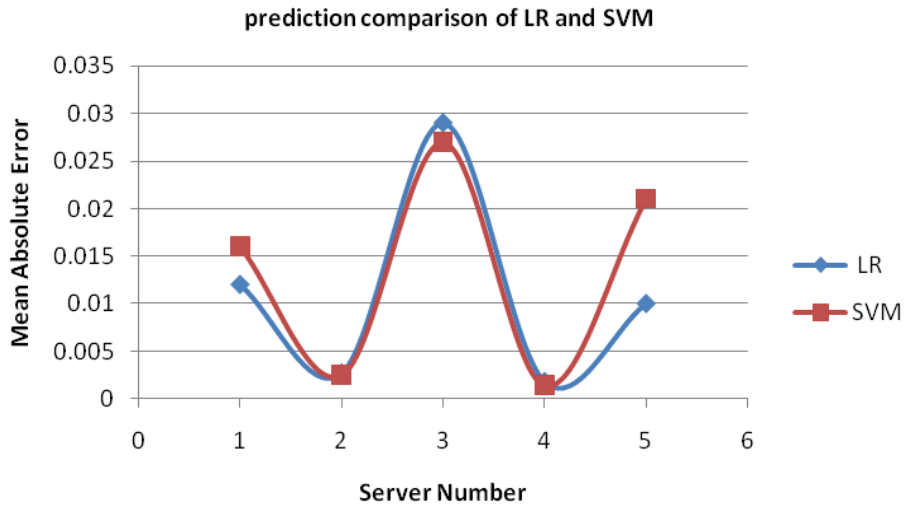

Figure 1 MAE comparison for various server logs

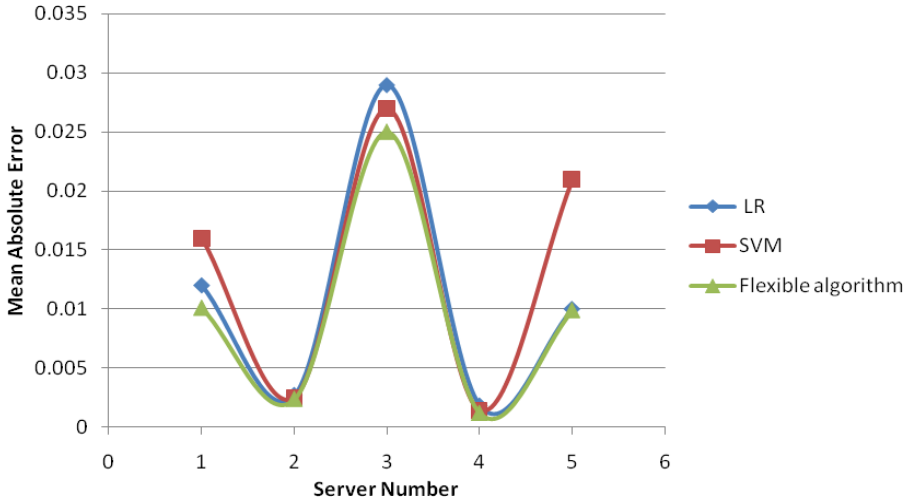

Figure 2 Comparison of the three methods

prediction error given by Linear Regression (meaLR) and prediction error of SVM (meaSVM). The minimum of the two has been assigned as the score of the algorithm. Finally, the average score of all the iterations has been outputted as the prediction error of the algorithm. We have tested the performance of the algorithm for the data set used earlier. Table IV presents the mean absolute error prediction of CPU utilization for data logs. Results clearly show that the proposed flexible algorithm performs better than the individual machine learning technique. Figure 2 illustrates the performance of the three methods used in the experiment.

TABLE II OBSERVED MAE FOR VARIOUS SERVER LOGS

S. No.	Server	MAE for LR	MAE for SVM
1	SANDIA-ROSS	0.012	0.016
2	LANL-O2K	0.0027	0.0025
3	LANL-CM5	0.029	0.027
4	HPC2N	0.0018	0.0014
5	LPC-EGEE	0.010	0.021

TABLE III THE FLEXIBLE ALGORITHM

```
1.  Load the data set
2.  Normalize data set
3.  Set n as the no. of folds
4.  For i = 1 to n
5.      calculate meaLR(i)
6.      calculate meaSVM(i)
7.      if meaLR(i) < meaSVM(i)
8.          set score(i) = meaLR(i)
9.      else
10.         Set score(i) = meaSVM(i)
11.  predicted_mea = score.mean()
12.  Return predicted_mea
```

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

TABLE IV PREDICTION ACCURACY OF THE THREE TECHNIQUES

S. No.	Server	LR	SVM	Flexible algorithm
1	SANDIA-ROSS	0.012	0.016	0.0101
2	LANL-O2K	0.0027	0.0025	0.0024
3	LANL-CM5	0.029	0.027	0.0025
4	HPC2N	0.0018	0.0014	0.0012
5	LPC-EGEE	0.010	0.021	0.0099

4. CONCLUSION AND FUTURE WORK

The load prediction method employed by the pro-active provisioning technique plays an important role in the overall efficiency of the provisioning method. Closer the future load predictions to the actual re-quirements, better is the performance of the provi-sioning technique in making the resources available to the applications as per the need. This paper evaluates the performance of Linear Regression and Support Vector Machine in predicting the CPU utilization of few servers. The prediction accuracy of the two tech-niques varies for different loads. The proposed flexi-ble algorithm combines the efficiency of the two techniques and provides better prediction accuracy. In future, we plan to use this prediction algorithm in designing a proactive provisioning strategy that can enhance the resource utilization in a cloud environ-ment.

REFERENCES

[1] P. Durgadevi and S. Srinivasan. (2017), "Resource discovery and dynamic resource allocation using MHACA and HOA algorithm in cloud environment", International Journal of Advances in Computer and Electronics Engineering, Vol. 2, Issue 1, pp. 11-15

[2] S. Singh and I. Chana. (2016), "Cloud resource provisioning: survey, status and future research directions", International Journal of Knowledge and Information System Vol. 49, Issue 3, pp. 1005- 1069.

[3] J. Zhang, H. Huang and X. Wang. (2016), "Resource provision algorithms in cloud computing: A survey", Journal of Network and Computer Applications, pp. 23-42.

[4] M. Amiri and L. Mohammad-Khanli. (2017), "Survey on prediction models of applications for resource provisioning in cloud," Journal of Network and Computer Applications Vol. 82, pp. 93-113.

[5] A. Matsunaga and J. A. B. Fortes, (2010), "On the use of machine learning to predict the time and resources consumed by applications," Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, Melbourne, Victoria, Australia, pp. 495-504.

[6] C. Vazquez, R. Krishnan, and E. John. (2015), "Time series forecasting of cloud data center workloads for dynamic resource provisioning", Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications, Vol. 6(3), pp. 87-110.

[7] H. Witten, E. Frank, and M. A. Hall, (2011), "Data Mining: Practical Machine Learning Tools and Techniques," 3rd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[8] J. Brownlee, "Machine Learning Algorithms from Scratch," ebook. Edition: v.1.5.

[9] J. Brownlee, "Machine Learning Mastery with Python," ebook. Edition: v.1.10.

[10] K. Cetinski and M. B. Juric. (2015), "AME-WPC: advanced model for efficient workload prediction in cloud," Journal of Network and Computer Applications, Vol. 55, pp. 191-201.

[11] N. Sapankevych and R. Sankar, (May 2009), "Time Series Prediction Using Support Vector Machines: A Survey," Computational Intelligence Magazine, IEEE, Vol. 4, Issue 2, pp. 24-38.

[12] Feitelson, D. G. Parallel Workload Archive, retrieved date: [2 February, 2018], online available at: http://www.cs.huji.ac.il/labs/parallel/workload/

[13] S. Islam, J. Keung, K. Lee, and A. Liu. (2012), "Empirical prediction models for adaptive resource provisioning in the cloud," Future Generation Computer Systems, Vol. 28(1), pp. 155-162.

[14] C. Liu, Y. Shang, L. Duan, S. Chen, C. Liu, and J. Chen, (2015), "Optimizing workload category for adaptive workload prediction in service clouds," Proceedings of the 13th International Conference on Service-Oriented Computing (ICSOC 2015), pp. 87-104, Goa, India, Springer-Verlag Berlin Heidelberg.

[15] Z. Chen, Y. Zhu, Y. Di, and S. Feng. (2015), "Self-adaptive prediction of cloud resource demands using ensemble model and subtractive-fuzzy clustering based fuzzy neural network", Journal of Computational Intelligence and Neuroscience. Vol. 2015, Article ID 919805, 14 pages.

[16] Montes J, Sánchez A and Pérez MS. (2011), "Grid global behavior prediction", Proceedings of 11th IEEE/ ACM international symposium on cluster, Cloud and grid computing, IEEE, Newport Beach, CA, pp. 124–133.

[17] L. Mao, D. Qi, W. Lin and C. Zhu. (April 2015), "A self-Adaptive prediction algorithm for cloud workloads", International Journal of Grid and High performance, pp. 65-76.

[18] Y. Jiang, C.S. Perng, T. Li, and R. N. Chang, (2013) "Cloud analytics for capacity planning and instant VM provisioning," IEEE Transaction on Network and Service Management, 10(3), pp. 312-325.

[19] T. Miu and P. Missier, (2012), "Predicting the execution time of workflow activities based on their input features," Proceedings of the SC Companion: High Performance Computing, Networking Storage and Analysis, SCC '12, Salt Lake City, UT, USA, IEEE Computer Society, pp. 64-72,

[20] S. Arlot and A. Celisse, (2010), "A survey of cross-validation procedures for model selection," Statistics Surveys 4, pp. 40–79.

[21] B. Efron and G. Gong, (1983), "A leisurely look at the bootstrap, the jackknife, and cross-validation," The American Statistician 37, pp. 36–48.

[22] Y. Dingyu, C. Jian, Y. Cheng, and X. Jing. (2012), "A multi-step-ahead CPU load prediction approach in distributed system," Proceedings of Second International Conference on Cloud and Green Computing, Xiangtan, Hunan, China, pp. 206-213.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 3, Issue 9, September 2018, pp. 7–11

Authors Biography

Padma D. Adane is Associate Professor at Shri Ramdeobaba College of Engineering and Management, Nagpur. She did her B.E. in Computer Technology and M.Tech in Computer Science from R. T. M. Nagpur University. She is currently pursuing her Ph.D. from Visvesvaraya Ntional Institiute of Technoogy, Nagpur. Her area of interest incudes Information Security, Mobile Computing and Cloud Computing.

Dr. O. G. Kakde is Professor and Head at Visvesvaraya National Institute of Technology, Nagpur and Ex-Director, VJTI, Mumbai. He did his M.Tech from IIT Bomaby and Ph.D. from Nagpur University at VNIT (formerly known as VRCE). He has a vast teaching and adminstrative experience. He has so far guided 4 Ph.D scholars with more than 25 research publications and authored 5 books. His area of interest includes Language Processor, Computer Programming Languages and Advanced Compiler Construction.

Cite this paper:

Padma D. Adane, Omprakash G. Kakde, "A Flexible Approach for Predicting CPU Utilization for Cloud Workloads", International Journal of Advances in Computer and Electronics Engineering, Vol. 3, No. 9, pp. 7-11, September 2018.