ONLINE RESOURCE USAGE PREDICTION AND FAILURE-AWARE SYSTEM FOR RESOURCE PROVISIONING IN CLOUD DATA CENTRES

A THESIS

submitted by

SHAIFU GUPTA

for the award of the degree

of

DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

MAY, 2020

Dedicated To

my mother and father for their endless love and support for me during toughest stages of my journey

THESIS CERTIFICATE

This is to certify that the thesis entitled **Online Resource Usage Prediction and Failure-Aware System for Resource Provisioning in Cloud Data centres** submitted by **Ms. Shaifu Gupta (Enroll. No: D14002)** to the Indian Institute of Technology, Mandi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her under my supervision. I recommend the work done in this thesis for the award of the Ph.D. degree. To the best of my knowledge, the contents of this thesis, in full or in part, have not been submitted to any other university or institute for the award of any degree or diploma.

Date:

Place:

Dr. Dileep A.D. (Supervisor)	Prof. Timothy A. Gonsalves (Co-Supervisor)
Associate Professor	Director and Professor
School of Computing & Electrical Engineering	School of Computing & Electrical Engineering
IIT Mandi, Kamand, India	IIT Mandi, Kamand, India

Acknowledgements

I would like to express my sincere gratitude to my advisors Dr. Dileep A.D. and Prof. Timothy A. Gonsalves for their continuous support and guidance during my Ph.D study. I am thankful to them for their motivation, patience, enthusiasm and immense knowledge. Their guidance helped me throughout my research and writing of thesis. I consider myself very fortunate for being able to work with very considerate and encouraging mentors like them.

I thank the members of my doctoral committee Dr. Samar Agnihotri, Dr. Padmanabhan Rajan and Dr. Manoj Thakur for their constructive feedback during my interactions with them. Their insight and questions were very helpful in streamlining the work.

I thank Dr. Aditya Nigam for his inputs and helpful suggestions during my work. I am thankful to Dr. Sriram Kailasam for insightful interactions about my work. I thank Dr. Anil Sao for his insightful comments and suggestions during our interactions. I thank my seniors Pulkit Sharma and Vinayak Abrol for the stimulated discussion and support.

I immensely express my heartiest thanks to all MANAS Lab faculty and students for encouraging the lively discussions and their support.

I thank Ministry of Human Resource Development (MHRD)- Govt. of India, for providing me stipend during my PhD. My sincere thanks to Indian Institute of Technology, Mandi for providing an environment to learn and grow. My special thanks to other faculty and staff members of IIT Mandi for providing an encouraging atmosphere all these years. I sincerely acknowledge the help and support of my staff friends, Nishant, Monika, Manoj and Lalit to name a few.

I also thank my friends for providing support when needed, keeping me motivated and giving me numerous moments to cherish in life. I thank Nivedita, Merlin, Swati, Akansha and Prabhjot.

Finally, this journey would not have been possible without the unconditional support from my mother Smt. Sapna Gupta. Her continuous urge to see me become successful in life has always motivated me to never give up. I would like to thank with my sincere and special gratitude my father Sh. Naval Gupta for always motivating me through my tough times.

Finally, thanks to God for the kind blessings.

(Shaifu Gupta)

ABSTRACT

A primary goal of cloud service administrators is effective resource provisioning. Prediction of future resource usage and proactive prediction of resource contention failures are two important functions of a resource provisioning system. Towards this goal, this work focuses on improving these two functions by analysing the nature of cloud resource usage workloads and by applying statistical and machine learning techniques.

The thesis first explores both experimental and analytical investigations to indicate the presence of long range dependence in cloud resource usage workloads. Thus, we compare the prediction capabilities of future resource usage prediction using models without and with modelling of long range dependence in resource usage workloads.

However, the performance of any resource may depend on other resources as well. Hence, this thesis proposes multivariate extensions of future resource usage prediction models. This work analyses six different multivariate frameworks based on regression models of all available resource metrics and a subset of relevant set of metrics. To support the selection of a relevant set of features, we propose to use two desirable characteristics of feature selection techniques, namely prediction performance and stability to support the dynamic cloud environment.

This thesis next proposes online variations for future resource usage prediction models. Here, the prediction models are updated in real time based on the error produced by the model. In this context, we analyse gradient descent and Levenberg-Marquardt parameter update methods. Further as long short term memory models for future resource usage prediction have a large number of trainable parameters, we propose sparse variations of these prediction models where only a few parameters are retained to support fast online adaptation of prediction models.

In addition to future resource usage prediction, this thesis also proposes a resource contention failure prediction system. Here, the heteroscedastic nature of cloud workloads is used with machine learning for anomaly detection. This work proposes an iterative autoencoder method for anomaly detection. Since detection of anomalies is not sufficient, we propose four different classification models for identification of types of anomalies into different kinds of resource bottlenecks. These include simple multiclass classifier, multiclass classifier with fractional differencing, multiclass classifier using encoded representation, and multiclass classifier using triplet-loss based representation.

To support continuous online learning of models used for identification of different types of anomalies, this thesis proposes a novel approach for real time adaptation of anomaly identification models. Here, we analyse two fundamental challenges associated with online adaptation of anomaly identification based classification models. These challenges are associated with catastrophic forgetting and architectural evolution in models. To avoid catastrophic forgetting, we propose a combination of standard loss and distillation loss in a teacher-student network approach. For architectural evolution, we propose three different alternatives for incremental architectural learning and a column subset selection based teacher-student network.

Based on the analysis, we observe that leveraging the presence of long range dependence, the proposed multivariate regression based prediction framework and the proposed online adaptation of prediction models, have enhanced the accuracy of CPU usage prediction by 69% over other existing methods for resource usage prediction. In the context of failure prediction, the proposed iterative autoencoder method performs 35% better than existing methods for anomaly detection. Among the proposed type of anomalies identification methods, LSTM-based multiclass classifier using triplet-loss based representation improves the performance of identification of types of anomalies by 66% which is further enhanced by proposed ensemble methods to a total of 76%.

Effective resource provisioning has become a necessity rather than a luxury. It is important that resource management decisions by schedulers and resource managers are carried based on the expected resource usage workload as well as the current state of the servers. Based on this, the work carried in this thesis achieves its primary goal to improve the performance of prediction of future resource usage and diagnosis of resource contention failures. Insights and outcomes from this thesis can be used by service administrators to achieve optimal resource management and its far-reaching impact on revenue, reliability and reputation of service administrators.

Keywords: Cloud data centre, resource usage prediction, failure prediction, long range dependence, neural networks, multivariate, online adaptation, sparse models, architectural evolution, long short term memory models.

Contents