Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmers Division
Second Semester 2018-2019

BITS ZG628T: Dissertation Midsem Report

ID No.	**: 2017HT12184**
NAME OF THE STUDENT	**:** Avijit Basu
EMAIL ADDRESS	**:** 2017ht12184@wilp.bits-pilani.ac.in
STUDENT'S EMPLOYING ORGANIZATION & LOCATION	**:** Juniper Networks, Bangalore
SUPERVISOR'S NAME	**:** Arun G Menon
SUPERVISOR'S EMPLOYING ORGANIZATION & LOCATION	**:** Juniper Networks, Bangalore
SUPERVISOR'S EMAIL ADDRESS	**:** agmenon@juniper.net
DISSERTATION TITLE	**:** Health Monitoring of VMs using Analytics

Health Monitoring of VMs using Analytics

BITS ZG628T: Dissertation

By

Avijit Basu

(2017HT12184)

Dissertation work carried out at

Juniper Networks, Bangalore

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE

PILANI (RAJASTHAN)

January-March 2019

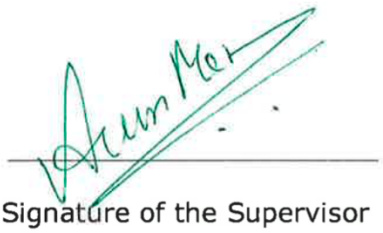
Dissertation Details:

1. Proposed topic of Dissertation: Health Monitoring of VMs using Analytics

Annexure C: *Format Certificate from the Supervisor*

CERTIFICATE

This is to certify that the Dissertation entitled Health Monitoring of VMs using Analytics and submitted by Avijit Basu having ID-No. **2017HT12184** for the partial fulfillment of the requirements of M.Tech. Software Systems degree of BITS, embodies the bonafide work done by him under my supervision.



Signature of the Supervisor

Arun G Menon
S/W Engineering Director
Juniper Networks.

Place : Bangalore

Date : 18/02/2019

Annexure D : Sample Format & Template of Abstract

Birla Institute of Technology & Science, Pilani

Work-Integrated Learning Programmes Division

Second Semester 2018-2019

BITS ZG628T: Dissertation

ABSTRACT

BITS ID No. : 2017HT12184

NAME OF THE STUDENT : AVIJIT BASU

EMAIL ADDRESS : 2017ht12184@wilp.bits-pilani.ac.in

STUDENT'S EMPLOYING ORGANIZATION & LOCATION : JUNIPER NETWORKS, BANGALORE

SUPERVISOR'S NAME : ARUN G MENON

SUPERVISOR'S EMPLOYING ORGANIZATION & LOCATION : Director **in** JUNIPER NETWORKS, BANGALORE

SUPERVISOR'S EMAIL ADDRESS: agmenon@juniper.net

DISSERTATION TITLE : Health Monitoring of VMs using Analytics

---

ABSTRACT :

Neural Networks are very good to classify and predict results from a given input set.
It can be shown that neural networks can implement universal logics like NAND and hence any logical functions can be fabricated using neural networks.

Neural networks are composed of a set of hyper parameters(viz number of neurons, number of layers) and a set of other parameters like weight of each connections, bias.
Accuracy of network is tuned by training the weights but hyper parameters cannot be trained.
Neural networks use gradient descent to minimize the loss functions and train weights. These works very well for smooth unimodal loss functions, but will converge at local minima for multi modal functions.

Genetic algorithm on other hand is robust to these above limitations. But can be very slow to converge. The idea is to use best of these 2 techniques.

Generate a population of neural networks, hyper parameters like activation function, loss measurement, number of neurons, number of layers as the genes. Then train each neural network on the input training set and estimate accuracy on the test set.
Then apply genetic techniques to evolve the population(accuracy being the measure of fitness), breed new neural networks by choosing hyper parameters from parent. Children can be muted too. Such a neural network can be used to predict VM loads based on VM parameters.

Broad Academic Area of Work: Machine learning

Key words

Genetic algorithms

Neural Networks

Virtualization

Signature of the Student

Name: AVIJIT BASU

Date: 18/02/2019

Place: Bangalore

Signature of the Supervisor

Name: ARUN G MENON

Date: 18/02/2019

Place: Bangalore

Acknowledgements

The satisfaction and euphoria that accompanies the successful completion of any task would be incomplete without mentoring the people who made it possible, because success is the epitome of hardwork, perseverance, undeterred missionary zeal, steadfast determination and most of all "Encouraging Guidance".

I express my gratitude to my supervisor Arun G Menon for providing me a means of attaining my most cherished goals.

I record my heart full of thanks and gratitude to my additional Examiner Ashish Kulshestra for providing me an opportunity to carry this project, along with purposeful guidance and moral support extended to me throughout the duration of the project work.

CHECKLIST OF ITEMS FOR THE FINAL DISSERTATION REPORT 18

List of Figures (with figure number, figure titles and page numbers)

List of Tables with table number, table title and page number.

Chapter 1:

Introduction

Cloud and virtualizations are a growing trend for Juniper Networks. A lots of Juniper routers can be deployed and host on cloud environment viz vMX, vPTX.
Also lots of internal IT resources like build etc are being host in contrail VM.
Growth and expansion of virtualized platforms emphasis the health monitoring of these VM hosts and quick turn around with proactive corrections.

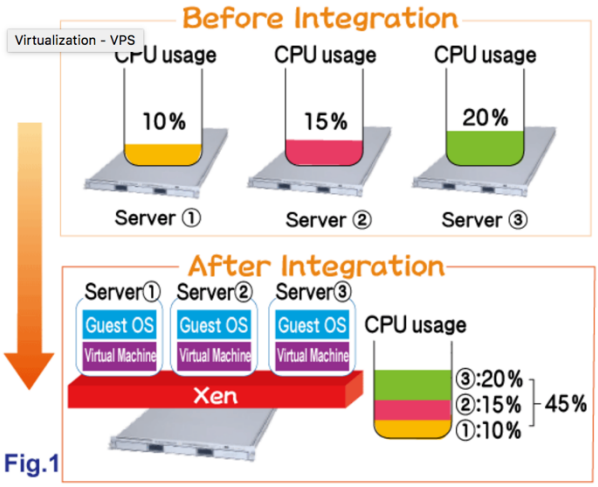
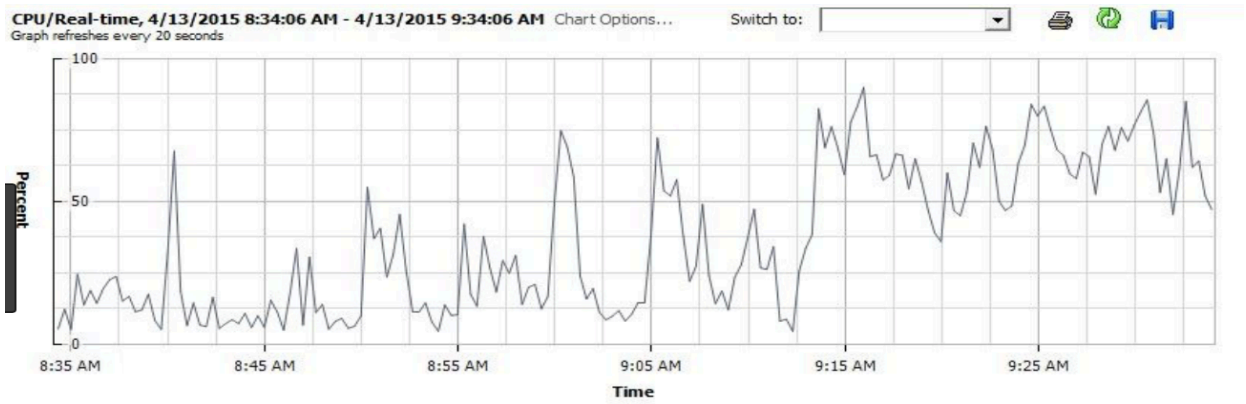


Figure 1: Virtualization

Analysis of health of VM/contrail involves understanding the correlation between various stats parameters like process profile, vm load, I/O operation stats. Automation is the key for quick detection of anomalies and then quick turn around for corrective actions. Automation with static rules does not reap much benefit. The rules need to be made dynamic.



Chapter 2

Approach To Problem

Monitoring of VMs involve the following

1. Automate a lots of scripts that collect a lots of data from VM clusters.
2. Look through the data and try to churn out following
    a. Is this VM really into resource crunch.
    b. If it is entering into danger zone, how long will this churn turn out to be.

Bigger Question: This VM cause churns to other VMs in this same Hypervisor. What will be future state of this VM.

Problem Statement

It is evident from Step1, there is hell lot of data to be analyzed and lots of prediction to be made.

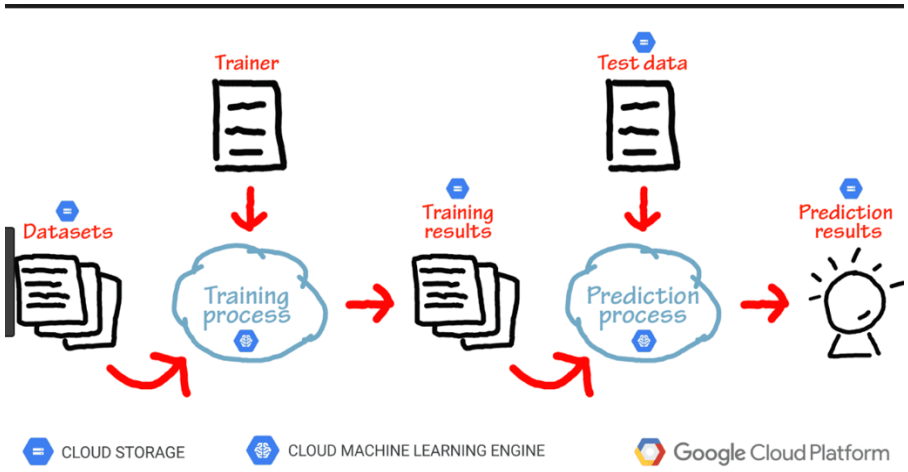
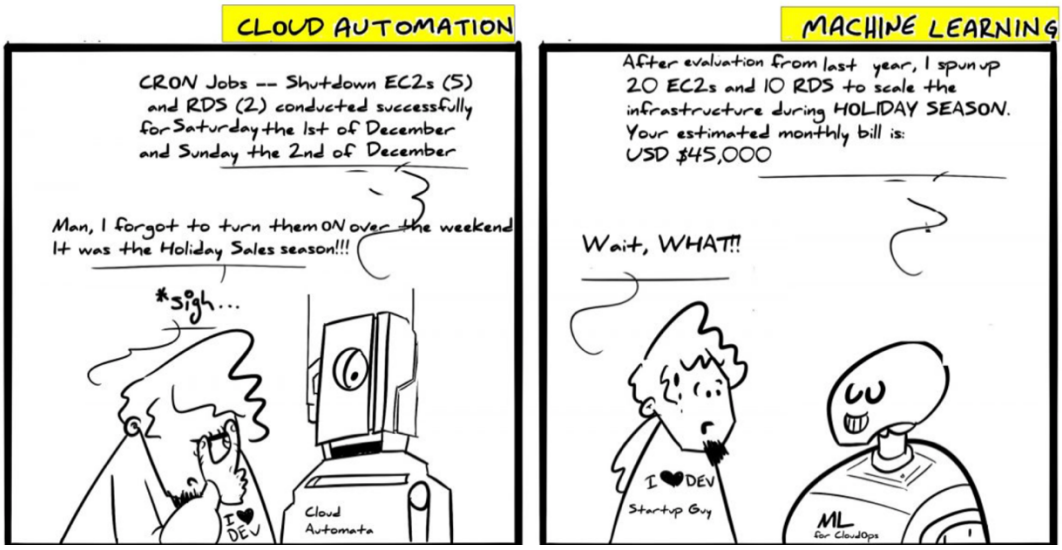


Figure 2: Generic Overview of Analytics

This calls for machine learning techniques like Neural networks. To used to analyze and predict results for VM hosts



Human vs Machine in terms of Efficiency !!!!

Neural Networks
What Neural Network in terms of Analytics.

Following picture depicts the concept of neural network.

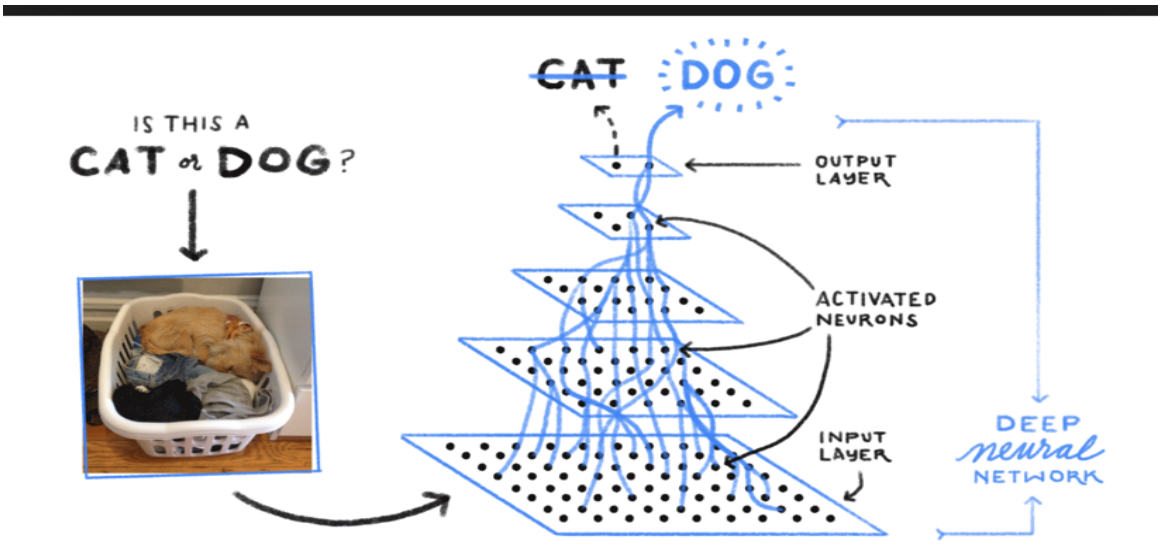


Figure 3: What is a Neural Network

Inputs to a Neural network triggers and activates a set of neurons. The activation ripples through one or more layer and finally one or more neurons at output layer.

So basically like any other algorithm, neural network maps an input set to an output set. The output can specify analysis or label on the input set. Output can be prediction of some future details too.
Ex the above networks can map a given picture into either a “Dog” or “Cat”

Neural networks differ from conventional algorithms

Conventional algorithms use cognitive approach. What to be done should be known and programmed into a set of instructions. Then for a given input it evaluates the instructions and then finds out what to execute next.

Neural networks cannot do miracles. But can be trained to produce output from a given input.

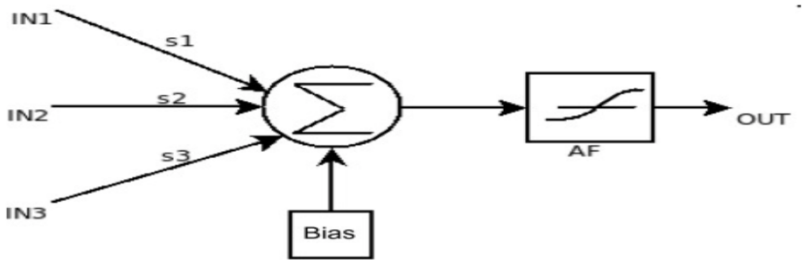
The mapping of input to output changes later, course of time.

No problem, we retrain the neural networks, as opposed to write a new algorithm in case of conventional algorithms.

How neural networks work

Neural network mostly use supervised learning.

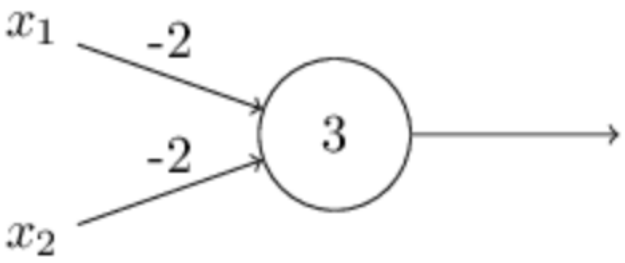
Each neuron is receiving a set of input either from the input data or activation from previous layer. Each of activation to the neuron has a weightage. Finally, the today weighted input is feed to an activation function of the neuron.



Multiplied by a weight matrix for each neuron input, the summation becomes a sum of weighted outputs from all other neurons and inputs. An activation function is then used to calculate individual neuron outputs.

What all can be implemented by a Neural network

Lets look at following network with 2 inputs X1 & X2. Each input has weights -2 and bias is 3.



X1	X2	Y
1	1	-1
1	0	1
0	1	1
0	0	1

Figure 4: NAND Logic with NN

If you look the above has evaluated an NAND logic.

It is well known any logic in computer computation can be evaluated using NAND(or NOR).

So this goes to prove neural networks can evaluate any logic.

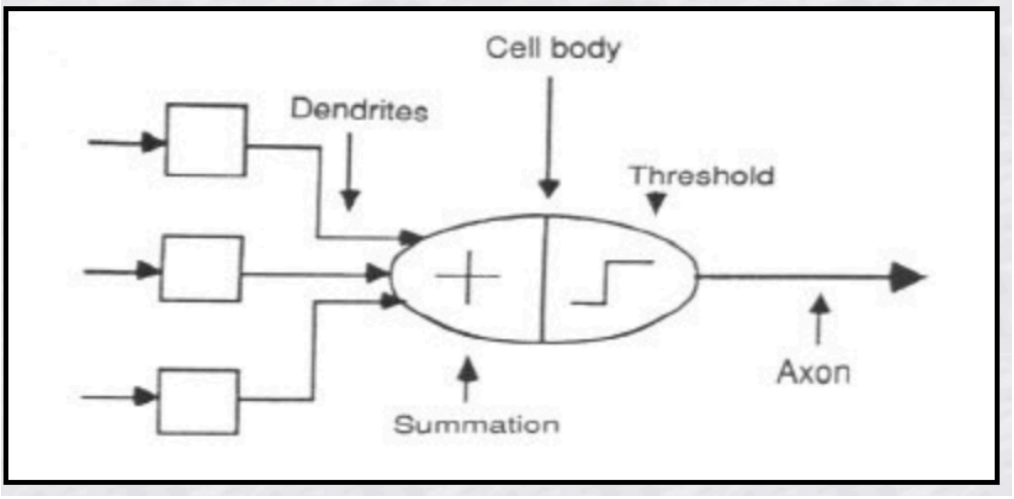
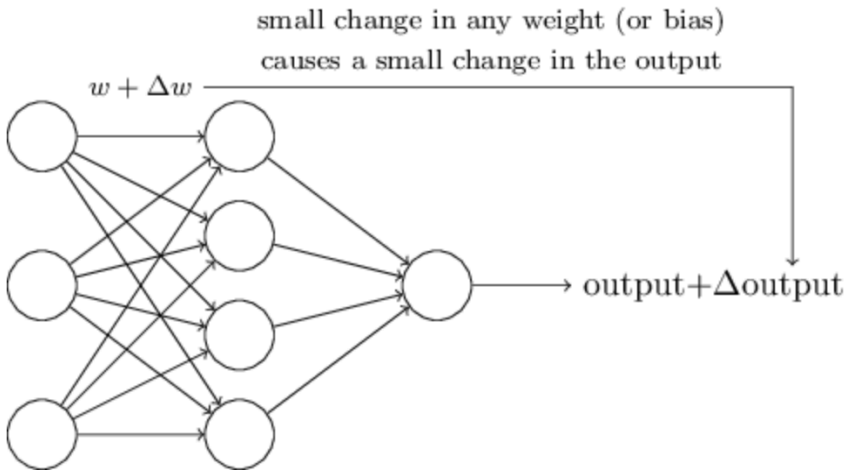


Figure 5: Analogy to Human neurons

Training neural networks

Training of neural network deals with adjusting the weight and bias of the individual connections, with a goal to minimize the loss of accuracy.



$$\Delta \text{output} \approx \sum_j \frac{\partial\, \text{output}}{\partial w_j} \Delta w_j + \frac{\partial\, \text{output}}{\partial b} \Delta b,$$

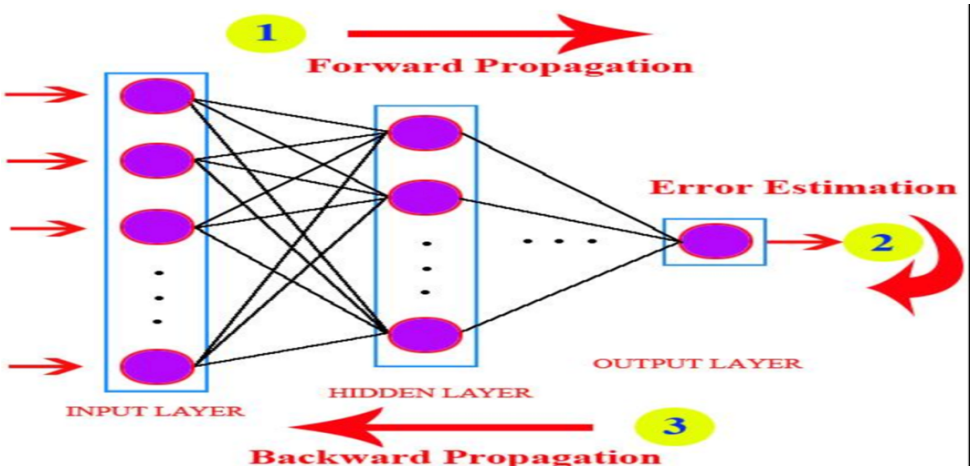


Figure 6: Minimizing loss by Backpropagation

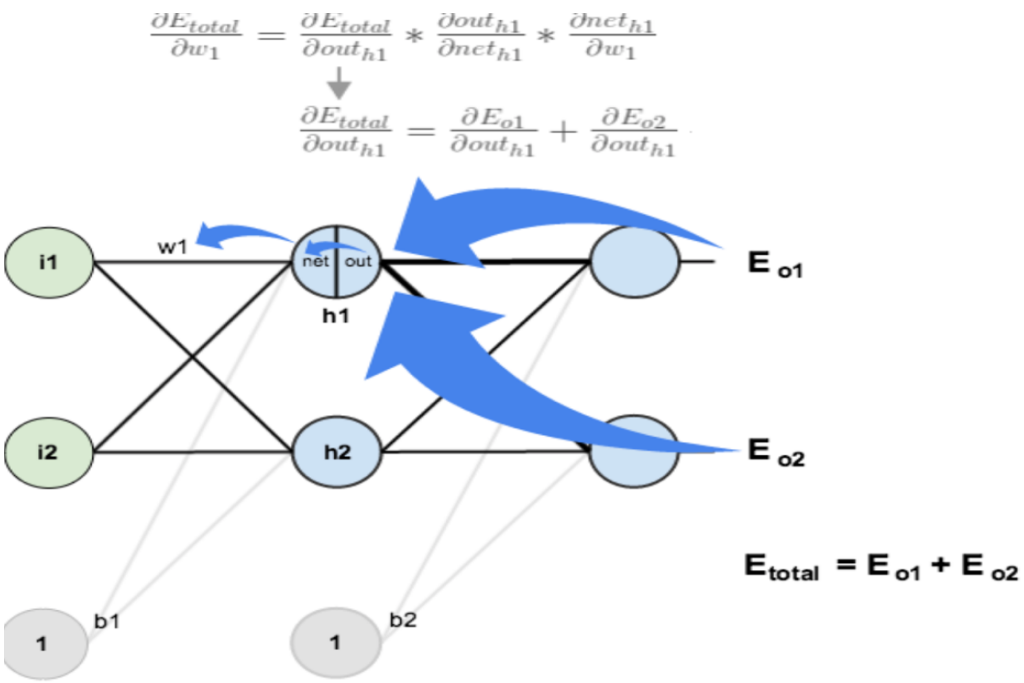


Figure 7: Adjustment of weight to minimize loss

Role of Neural Network in this Algorithm

Proposed ML model to tackle the problem:

- A neural net trained via supervised learning
- Supervised learning is achieved by massaging the VM(s) history data as follows:
- Take an average of first 10 rows of VM data.
- If every row represents data captured/sec, then we take average of first 10 secs of VM data.
- Based on the resource utilisation computation which factors in CPU usage, memory access, incoming and outgoing jobs, we classify the dataset into very hot, hot, warm, cold and very cold, visualised as below,

Input to Neural network(Snippet of DataSet):

1. CPU usage percentage
2. Memory access
3. I/O rate
4. Incoming job rate
5. Out going job rate
6. Processes with /bin/sh /volume/build pattern count (Parallel builds)
7. 'Processes with "bmake" pattern count'
8. 'Processes with "emake" pattern count'

Table 1: Data Set(Snippet)

Class number	**Resource utilisation (formula)**
1(very cold)	0-20%
2(cold)	20-40%
3(warm)	40-60%
4(hot)	60-80%
5(very hot)	80-100%

Table 2: Classification of VM based on load

Prediction:

For a virtual machine, factoring in the history of utilization of computational resources on or before time t1, we try to predict the CPU utilization at time t2 where
t2 = t2+x {x / x!=0, x->Z; 1<= x <= # of rows used for resource utilisation computation}.

Results:

Confusion matrix is used to describe the performance of our classification model, as the solution inclines to identifying in which Class number, the VM ends up in future time.

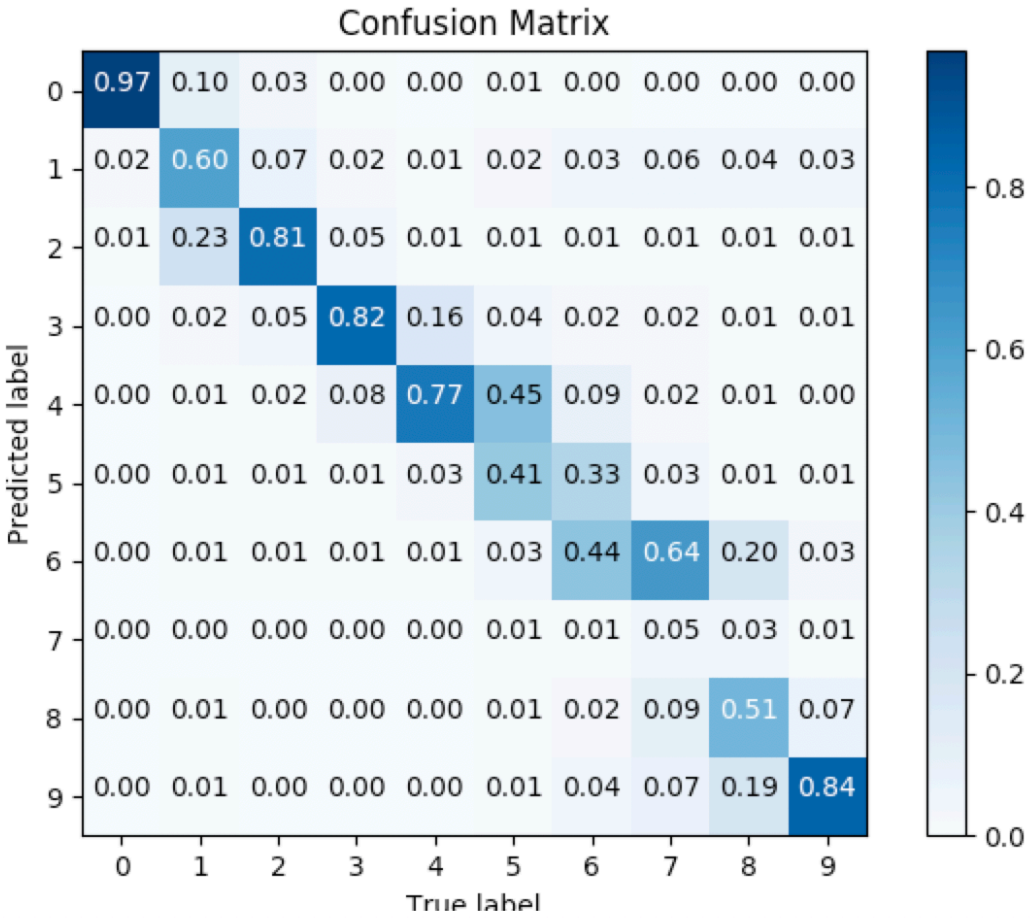


Table 3: Confusion Matrix(to evaluate NN)

Now Neural Networks have limitation too

NNs have helped us solve so many problems. But there's a huge problem that they still have. **Hyperparameters!**

Hyperparameters are the following:

1. Number of neurons in the network
2. Activation function
3. Loss optimizer(used while training)

These are the only values that can not be learned… Until now.

If we choose an improper set of thee hyperparameters, it will take long time and cycles of training before we improve accuracy.

Augment Genetic Algorithms to Neural networks.

A genetic algorithm is as accurate a model of evolution as an artificial neural network is a model of the brain.

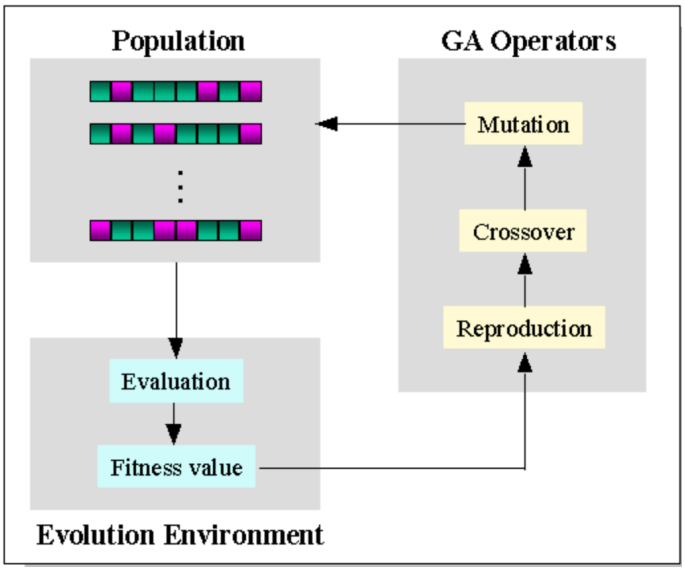


Figure 8: How Genetic Algo work

Take the Best of Neural Networks and Genetic Algorithm and combine them.

Generate a population of neural networks with a specific set of hyper parameters.

Then train each network and generate the accuracy.

Choose the set of fittest among them and produce a set of children from these fit parent.

Child will can random hyperparmeters from parent.

Train the new population.

This is the Algo being used in our monitoring software.

Neural networks will be used to predict state of VM, inputs being monitoring data.
We choose hyperparameters of the networks using Genetic algorithms.

Over view of Algorithm

This algorithm relies on Neural Network for training and Genetic algorithm for evolving a population of neural networks and choosing the fittest.

Note*: There will be some work required for preprocessing of the Data set. Those details are being worked out and will be present in the next set of reports.*
This report mainly emphasis on the Neural network and Genetic evolution.

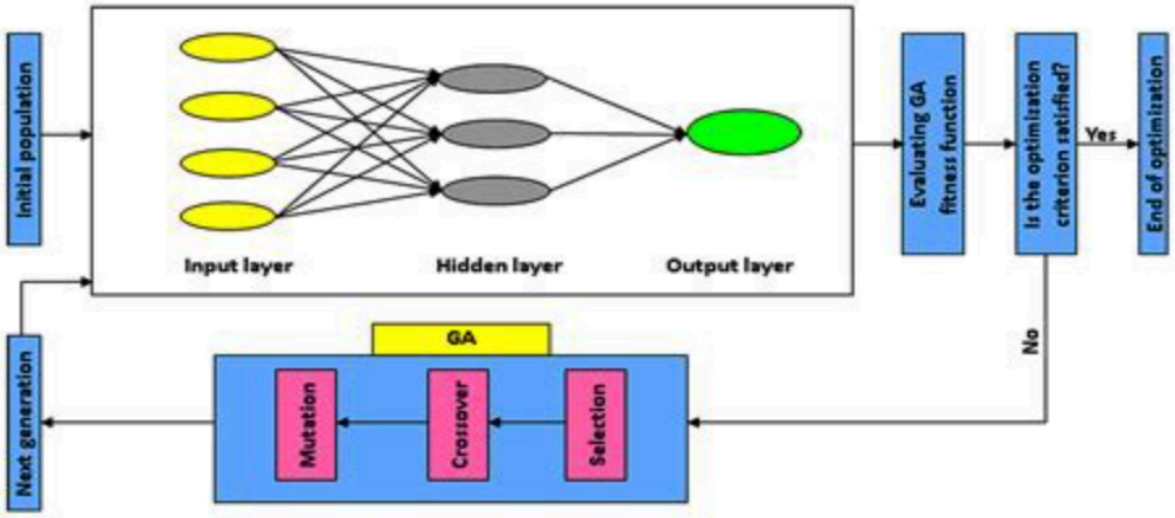


Figure 9: Algorithm used for this project

Steps of Algorithm:

1. Create a set of Neural networks by choosing randomly from a set of hyper parameters.
2. Train each Neural network based on the training set chosen for the data collected from the VMs.
3. Using the testing set, assign accuracy to each NN.
4. Sort NNs based on accuracy.
5. Keep the fittest set as parents. Accuracy on the training set, is the fitness function chosen.
6. Randomly choose a few parents and generate child.
    a. Child takes hyperparameters from each of the 2 chosen parent.
    b. For each hyper parameter of child take randomly from hyper parameters from each of 2 parents.
    c. This is where I need to experiment a bit and maybe tune the logic, like choose certain parameters from the more fitter parent.
7. New population comprises of the fittest parent from previous generation and new set of child.
8. Rerun training on this population.

    Few things to note here.

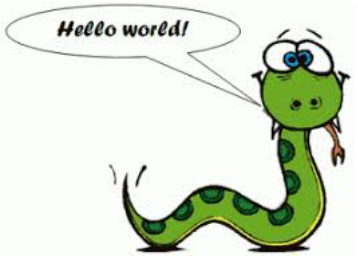
    a. Parent already have some weights from previous training, start from that. This reduces training time of parent.
    b. Childs off course take start training with random(non zero weight).
9. Run the above for a few generations and then choose the top few neural networks.

Result

Confusion matrix is used to describe the performance of our classification model, as the solution inclines to identifying in which Class number, the VM ends up in future time.

Chapter 4

Software Details



Data Science in Python

This Software being developed in Python.

Software has following parts

1. Code to generate a population of neural networks
2. Code to train each neural networks on the training set
3. Code to estimate accuracy of neural networks on the testing set.
4. Code to sort the population based on accuracy.
5. Code to breed child neural networks from fittest parents.
6. Retrain the fittest parents and children on the dataset. This is done generation over generation.
7. Finally after some generation choose the top few networks.
8. Code to preprocess the input data set. This is to sort out outlier, etc. This will help achieve training accuracy.

Activities completed so far:

Activity 1, 4, 6 are complete.
To test logic of genetic algorithm, accuracy of each neural network updated by some random number generator.

Code for neural network training (activity 2, 3) are in progress.
Activity 8 will be done.

Status of Work done

In reference to outline document

Date	Work Plan	Status
11/01/2019- 01/02/2019	Evaluation of the model and algorithm of the software. This includes determine what ML techniques to be used.	Completed, Finalized to use Neural networks. Augment genetic algorithm to tune hyper parameters
02/02/2019- 15/02/2019	Design of finer and lower level details.	Completed: Algorithm has been stated.
16/02/2019- 10/03/2019	Implementation of code and attempt to refine the algorithm to improve accuracy.	In progress
11/03/2019- 31/03/2019	Testing, Final Report Submit	Yes to be started

Table 4: Progress Report of Project

Summary

Investigate efficiency of neural networks to predict future CPU usage of specific VM(s) under study based on history of the VM(s) resource utilization.

Confusion matrix is used to describe the performance of our classification model, as the solution inclines to identifying in which Class number, the VM ends up in future time.

Hyperparameters of Neural networks can be improved and tuned through genetic algorithm.

Conclusions and Recommendations (if any)

The load prediction can be further input to VM maintenance algorithm responsible for VM provisioning. But that is out of scope for this project.

Directions for future work

Right now the genetic algorithm to generate and evolve a population of neural networks have been put in place.
Next key step is to train individual networks and generate a population of neural networks of high accuracy to the genetic algorithm.
One other aspect that be explored going forward, feeding to this neural network algorithm, more set of monitoring parameters of VM.

Bibliography (if any) (Please refer to the sample format given below)

http://ijarcsse.com/Before_August_2017/docs/papers/Volume_6/8_August2016/V6I8-0121.pdf

Checklist of items for the Final Dissertation Report

This checklist is to be attached as the last page of the report.

This checklist is to be duly completed, verified and signed by the student.

1.	**Is the final report neatly formatted with all the elements required for a technical Report?**	Yes
2.	Is the Cover page in proper format as given in Annexure A?	Yes
3.	Is the Title page (Inner cover page) in proper format?	Yes
4.	(a) Is the Certificate from the Supervisor in proper format? (b) Has it been signed by the Supervisor?	Yes Yes
5.	Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly?	Yes Yes
6.	Is the title of your report appropriate? **The title should be adequately descriptive, precise and must reflect scope of the actual work done.** Uncommon abbreviations / Acronyms should not be used in the title	Yes
7.	Have you included the List of abbreviations / Acronyms?	Yes
8.	Does the Report contain a summary of the literature survey?	No
9.	Does the Table of Contents include page numbers? (i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1) (ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures) (iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables) (iv). Are the Captions for the Figures and Tables proper? (v). Are the Appendices numbered properly? Are their titles appropriate	Yes Yes Yes Yes Yes Yes
10.	Is the conclusion of the Report based on discussion of the work?	Yes
11.	Are References or Bibliography given at the end of the Report? Have the References been cited properly inside the text of the Report? Are all the references cited in the body of the report	Yes Yes Yes
12.	Is the report format and content according to the guidelines? The report should not be a mere printout of a Power Point Presentation, or a user manual. Source code of software need not be included in the report.	Yes

Declaration by Student:

I certify that I have properly verified all the items in this checklist and ensure that the report is in proper format as specified in the course handout.

Avijit Basu

Place: BANGALORE — **Signature of the Student**

Date: 18/02/2019 — **Name: AVIJIT BASU**

ID No.: **2017ht12184**