Test Lines with Common Topology & Dynamic Allocation

Tarun Bharti
Principal Technical Architect,
TechMahindra,Electronics City
Bangalore, India
TB00846057@techmahindra.com

***Abstract*— In CI/CD we need to always ensure green trunk & faster deliveries, In the DevOps model, testing is done continuously,We need to maintain multiple test topologies for multiple product variants.Apart from trunk we have multiple branches where CI runs in parallel.In Current CI environment we have One To One mapping between test case & test lines,as more and more branches increases in future we need to replicate the same test topology multiple times,which is not cost effective in terms of Setup Maintenance cost Opex,Real Estate Utilization,Power Consumption and Capex involved related to lab inventories.**

In this whitepaper we propose the use of common test topology with dynamic allocation of testlines , where any test case can be executed on any testline available in the pool dynamically.Due to dynamic nature of testline allocation it improves the testline utilization & due to common test topology any test scenarios can be tested on any testline , N to One & we have a fallback to test the coverage in other available testline in case of any testline related failures .In this way we will have to maintain lesser number of redundant setups and will also ensure high availability of testlines.

Keywords –Dynamic allocation,Common topology , CI/CD,DevOps

INTRODUCTION

CI RCA has become a major part of Software industry. All the failure and actions are usually tracked down by CI RCA tool. Assignment of these RCA to designated user or user group become a major and key task. Many times observed that lot of failures are not updated timely on RCA tool and due to this advancement there are large amount of unprocessed or partially process information present in RCA tools. In such cases it becomes difficult for RCA owner to get more details of failure just by looking at it. Also it is time consuming to view or extract the needed information from CI execution logs. In such a situation we are in need to develop a strategy which is useful to obtain the necessary information and update it to RCA tool and assign the RCA to particular user or user group. Since there are large amount of data decision manually doing it is very tedious. To overcome these pitfalls the concept of Classification using Machine Learning is used. Classification is the process of filtering relevant data according to one's business interests from the huge collection of data using different techniques and algorithms.

PROBLEM STATEMENT

The main purpose of the tool being developed is to enable automatic assignment of RCA with proper failure logs to corresponding teams. This assignement must be accurate and the same shall be reflected back to the CI RCA tool. Also, the model for classification must be trainable and easy to use for any other field corresponding to RCA , if required in future.

SOLUTION

As a solution we propose the mechanism as shown in Figure1.

1. One of the first step is to fetch all available data about CI Failure. As of now all the CI Failure logs are already available in RCA Tool. In order to get all required information about CI failure we will have to make a REST API call to RCA Tool. With the API call, all necessary infromation about the failure will be fetched. All these fetched information will be stored locally in our local database.
2. Once we have all the available data, we need to use relevant classification models to classify. We will be splitting our data into 2: Test data & Training data. The training data will be used to train and generate the models
3. After training the classififer a model will be generated. This model will be evaluated for required accuracy cut-off. If the model passes the accuracy test, the model can be used for classification of RCA.
4. Test with real data. Once we have the model, we will use the model for automatic RCA team assignement using the generated model

CLASSIFICATION

In machine learning and statistics, classification is the technique of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient. Classification is the process of converting the data records into set of classes.

It is divided into two:

- Supervised classification
- Unsupervised classification.

In supervised classification, the data that are to be classified is previously known based on few assumptions. In Unsupervised classification, the set of cases were not predicted by the users.

By some assumption it is the job of the user to classify the given data and try to assign the name for those cases. Classification involves predicting a certain outcome based on a given input. In order to predict the results, it needs to fetch the data already available. Based on this data the records are classified. The data sources can be categorized into training set and test set. The training set contains the data which are classified before and it used as a reference for classification purpose. With the help of the attributes the results are predicted.

The algorithm analyses the data given and predicts the results. Some of the algorithm against which the RCA assignement has been tested include:

1. Random Forest Classifier: It is an ensemble tree-based learning algorithm. Ensemble algorithms are those

which combines more than one algorithms of same or different kind for classifying objects. example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. A decision tree is widely used classification technique. The methodology used here is Divide and conquer. As there were huge amount of data, first we need to divide those data into sub data. The structure of the decision tree is organized in a manner that it contains the root the topmost node in the tree, Branches which are the internal nodes and leaf node is one which is not further classified. The internal nodes represent a question and the branch which connects the node denotes the solution and the leaf node tries to predict the solution.

2. MultiLayer Perceptron Classifier: In the Multilayer perceptron, there can more than one linear layer. If we take the simple example the three-layer network, first layer will be the input layer and last will be output *layer* and middle layer will be called *hidden layer*. We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task.

3. Stochastic Gradient Descent Classifier: The word '*stochastic*' means a system or a process that is linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In Gradient Descent, there is a term called "batch" which denotes the total number of samples from a dataset that is used for calculating the gradient for each iteration. In SGD, it uses only a single sample, i.e., a batch size of one, to perform each iteration. The sample is randomly shuffled and selected for performing the iteration.

The automatic assignment of RCA have been done using these 3 indepenedent algorithms. The accuracy of each of the algorithm has been tested and tabulated as seen in Table1.

FIGURES & TABLES

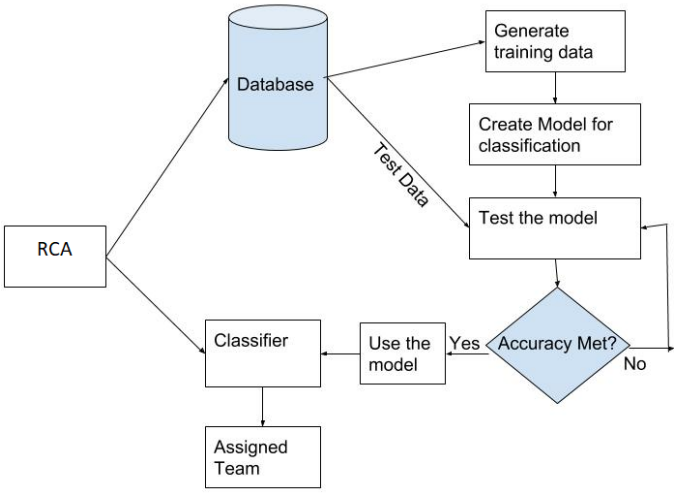

Fig 1. Overview of RCA Assignment system

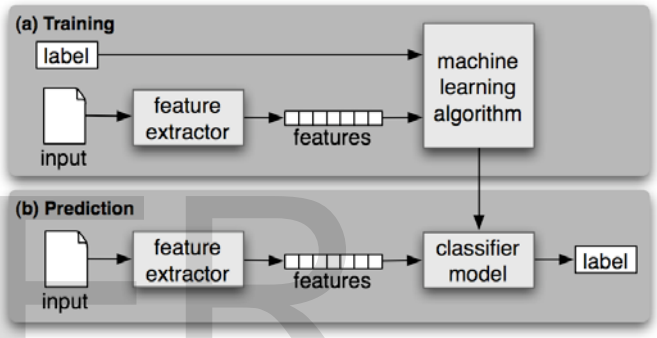

Fig 2. Supervised model of classification

Table 1. Accuracy of Different classifier

	Random Forest Classifier	MultiLayer Perceptron Classifier	Stochastic Gradient Descent Classifier
Testing Accuracy	58.86%	62.56%	52.12%
Training Accuracy	96%	93 %	93%

Table 2. Training & Test Data

	Cluster\Area	RcaWhys	Jira Action
Training data	1251	914	943
Test Data	1242	908	423

Note: Above data is represented no. of CI Job Cycle ID's

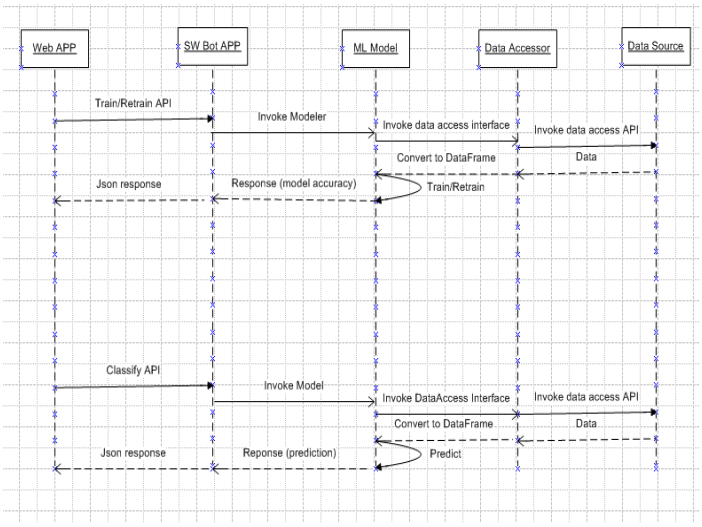

Fig 3. Sequence diagram for RCA Assignment

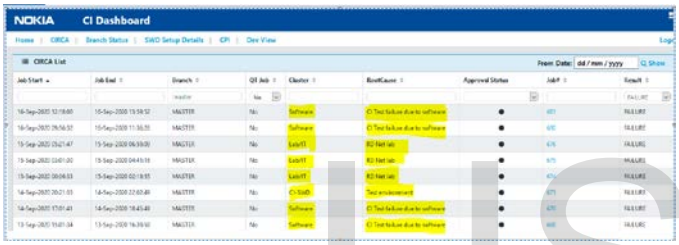

Fig 4. CICD Dashboard for CI Failure

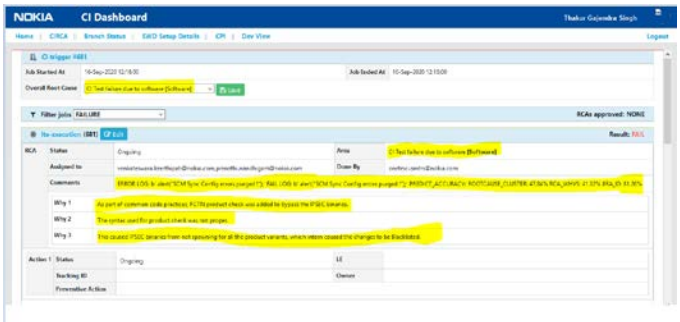

Fig 5. Auto RCA done by Machine Learrning

Trainingdata

Sample: Training Data

CONCLUSION

This paper gives the basic idea about using classification techniques for Automatic RCA assignment based on failure Error Logs using Random Forest Classifier, MultiLayer Perceptron Classifier, Stochastic Gradient Descent Classifier and the accuracy using the same have been mentioned. These algorithm performances can also be evaluated using the criteria Sensitivity, Accuracy, Error Rate, precision, recall and f-measure. Its observed the the MultiLayer Perceptron Classifier has a better accuracy rate than the remaining other algorithms and is more suited for automated assignment of RCA.

REFERENCES

[1] Aarti Sharma Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, "Application of Data Mining: A Survey Paper", International Journal of Computer Science and Information Technologies, Vol. 5(2), 2014, 2023-2025 ISSN:0975-9646.

[2] S.Neelamegam,Dr.E.Dharmaraj, "Classification Algorithm in Datamining:An Overview", International Journal of P2P Network trends & technology, Vol.4,issue 8-Sep 2013.

[3] George Dimitoglou, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability".

[4] Fiona Nielsen,"Neural Networks-Algorithms and Applications".

[5] Dr.S.Vijayarani,Mr.S.Dhayanand,"Datamining Classification Algorithm for Kidney Disease Prediction", International Journal on Cybernetics & Informatics, Vol 4,Nov 4,Aug 2015.