INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue1)
Available online at: www.ijariit.com

An Intelligent Scientific Workflows Failure Prediction Model Using Ensemble Learning Technique

Parminderjeet Kaur
Research Scholar
CSE Department, Punjabi University
Patiala
Parminderjeetkaur73@gmail.com

ABSTRACT: Cloud computing is a distributed computing paradigm which is considered as the computing platform that is going to be the pioneering field for the next ten years. Apart from several industrial, business applications being deployed, this paradigm is additionally attracting several scientific communities to utilize the services of cloud for running massive scale knowledge and computation intensive applications like a montage that is employed in astronomy. Workflow is defined as a group of task and dependencies between the tasks that are used for expressing numerous scientific applications. The main issue in running these workflow applications is mapping the tasks of the workflow to an appropriate resource in the cloud environment. Scheduling these workflows in a computing environment. To overcome these failures, the workflow scheduling system should be fault tolerant. The fault tolerance by using replication and resubmission of tasks supported the priority of the tasks. The replication of tasks depends on a heuristic metric that is calculated by finding the trade-off between the replication issue and resubmission issue. As scientific workflows scale too many thousands of distinct tasks, failures because of the software package and hardware faults become progressively common. We study job failure models for data collected from different scientific applications, by our proposed framework. In particular, we show that the Ensemble learning classifier can accurately predict the failure probability of jobs. Failure prediction models have been implemented through machine learning approaches and evaluated performance metrics. The models allow us to predict job failures for a given execution resource and then use these failure predictions for two higher-level goals: (1) to suggest a better job assignment, and (2) to provide quantitative feedback to the workflow component developer about the robustness of their application codes.

KEYWORDS:Cloud Computing; scheduling; Ensemble Learning, fault tolerance.

I.INTRODUCTION

Cloud computing is a kind of service based on internet which allows the user to access any number of resources dynamically on demand from anywhere and anytime in a metered manner i.e. pay per usage without paying much heed to the maintenance and implementation details of application. It is based on three computing concepts: Grid computing, utility computing and Autonomic computing [11]. Grid computing is a service provisioning heterogeneous distributed system that is used for executing the highly complex applications requiring huge processing power and massive amount of data. The renting of computing resources like network bandwidth, hardware and software on demand as per the requirement is known as utility computing. Autonomic computing means capable of self-management. Cloud computing is the combination of loosely coupled heterogeneous devices which are connected via internet providing the wide range of services according to the specifications mentioned in the service level agreement. Cloud has become a commercial commodity hence it is very important to consider the various QoS parameters like cost, time, reliability, and security while offering the services to clients to satisfy the customers. It is a prominent technology which provides both hardware (Amazon EC2) and software services (Google apps) along with the storage facility as per user's request dynamically over the internet using the pay per usage model. It has the efficiency to handle large loads with less cost investment.

The main issue in running these workflow applications is mapping the tasks of the workflow to an appropriate resource in the cloud environment. Scheduling these workflows in a computing environment. To overcome these failures, the workflow scheduling system should be fault tolerant.

II. EXISTING WORK

Gong, C., et al. (2010) proposed Low Latency Fault Tolerance (LLFT) Model that utilizes leader/follower replication approach and provides fault tolerance for distributed applications deployed within a cloud computing environment. The novel commitments of the LLFT middleware incorporate the low Latency Messaging Protocol, the leader-determined membership protocol and the virtual determinate Framework. [1]

Sun, D. W., et al. (2012) put forward a dynamic adaptive fault tolerance strategy (DAFT) that is focused around the standards and semantics of cloud fault tolerance. An analysis on relationship between different failure rates and two different fault tolerance techniques, check-pointing and replication has been carried out. A dynamic adaptive model has been built by combining the two fault tolerance models which helps to increase the serviceability. [2]

Bala, A et al. (2014) put forward an idea of designing an intelligent task failure detection models for facilitating proactive fault tolerance by predicting task failures scientific workflow applications. The working of model is distributed in two modules. In first module task failures are predicted with machine learning approaches and in second module the actual failures are located after executing workflow execution in cloud test-bed. Machine learning approaches such as naïve Bayes, ANN, logistic regression and random forest are implemented to predict the task failures intelligently from the dataset of scientific workflows. [3]

Xiong, N., et al. (2012) Given that networks are dynamic and unexpected, Naixue-Xiong, investigates Failure detector properties with connection to real and programmed fault-tolerant cloud based network systems, in order to discover a general non-manual investigation strategy to self-tune corresponding parameters to fulfill user requirements. Based on this general self-tuning method, they propose a dynamic and programmed Self Tuning Failure Detector scheme, called SFD, as an improvement over existing schemes. [4]

Meshram, A et al. (2013) proposed fault tolerance model for cloud (FTMC). This model accesses the reliability of computing nodes and choses the node for the computation on the basis of reliability. The node can be removed if it does not perform well. [5]

Jhawar, R et al. (2013) provided a new dimension for applications deployed in a cloud computing infrastructure which can obtain required fault tolerance properties from a third party. The model straightforwardly works fault tolerance solution to user's applications by combining selective fault tolerance mechanisms and discovers the properties of a fault tolerance solution by method of runtime monitoring. [6]

Joshi, S et al (2014) proposed a fault tolerance mechanism to handle server failures by migrating the virtual machines hosted on the failed server to a new location. Virtualization has been applied for data centers giving rise to the concept of virtual Data Centers (VDC) which have virtual Machine (VM) as the basic unit of allocation. Using appropriate resource allocation algorithms, multiple VDCs can be hosted on a physical data center. [7]

Huang, S. et al (2010) proposed Dual Agreement Protocol of Cloud Computing (DAPCC), keeping in consideration the scalable and virtual nature of cloud. DAPCC is proposed to tackle the agreement problem caused by faulty nodes which send wrong messages, it tells how the system achieves agreement in a cloud computing environment. [8]

Nguyen, H (2013) proposes that one of the biggest challenges for diagnosing an abnormal distributed application is to pinpoint the faulty components. Black-Box online fault localization system called F-chain has been presented that can pinpoint faulty components immediately after a performance anomaly is detected. F-chain is presented as: a practical online fault localization system for large scale Iaas clouds. This system does not depend upon prior knowledge i.e. previously seen and unseen anomalies, and is practical for Iaas clouds. To achieve higher pinpointing accuracy, an integrated fault localization scheme has been introduced that consider both fault propagation patterns and inter component dependencies. [9]

Lima, F et al (2004) proposed adaptive failure detectors that are adjustable to the changing communication loads and use artificial neural networks for predicting the arrival time of next heartbeat from a virtual machine. [10]

III. EXISTING PROBLEM

Accurate failure predictions can help in mitigating the impact of failures for scientific applications and resources, applications, and services can be scheduled efficiently to edge the effect of failures. However, providing accurate predictions sufficiently ahead is a challenging task for intricate applications such as workflows and an accurate prediction of task failure is a pre-requisite for implementing intelligent fault tolerant approach for scientific workflows. The state of the art of existing approaches for failure prediction revealed that some of the research challenges can be resolved using intelligent failure prediction models on large infrastructure such as Cloud. In the existing work, some machine learning algorithms are used to predict the faults in scientific applications and show the naïve bayes algorithm results to be better than the other algorithms. But naïve bayes algorithm assumes independency of features that leads to less accurate prediction model. Another problem happens due to data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequents approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results.

IV.PROPOSED WORK

1. Collection of Dataset

Collection of failure data from Montage, Cybershake, Inspiral and Sipht workflow applications using Workflow Sim.

2. Preprocessing and Filtration

Preprocessing and filtration of the collected data. Filters like Replace Missing Values and Numeric to Nominal filter.

3. Failure Prediction using Machine Learning

Classification is a technique for machine learning by which it is used to predict the grouping membership of different data instances. It will perform the task by which it will generalize the well-known structure so as to apply it on new data. Here ensemble classifier has been used for quality measurement of dataset will be consider on the basis of percentage of correctly classified instances.

In the first phase each of the base level classifiers takes part in the j- fold cross validation training where a vector is returned in the form < (y'0… y'm), yj > where y'm is the predicted output of the mth classifier and yj is the expected output for the same

In the second phase this input is given for the Meta learning algorithm which adjusts the errors in such a way that the classification of the combined model is optimized. This process is repeated for k-fold cross validation to get the final combined generalization model.

- For first Phase, AdaBoost algorithm is used and acts as a base leaner
- In Second phase, Decision Tree algorithm is used as a Meta classifier.

4. Comparing the results of the proposed machine learning technique with the previous techniques on the basis of above mentioned parameters
5. Also, the prediction of actual failures with the predicted failures is done.

V. SIMULATIONS RESULTS

The proposed methodology is implemented with the help of Work flow Sim and Net beans IDE 8.0. Work flow Sim is the library that provides the simulation environment of cloud computing and also provide primary classes describing virtual machines, data centers, users and applications. The performance has been evaluated on the basis of following parameters:

- **Accuracy:** Accuracy is percentage of testing set examples correctly classified by the classifier. It is the proportion of total number of predictions that are correctly classified in class.
- **Precision:** Percentage of selected instances that are relevant and are correctly classified in class out of all documents in class.
- **Recall:** Percentage of correct documents that are selected in class from the entire document actually belonging to class.
- **Sensitivity and specificity**: These metrics measures the correctness of the predicted model where Sensitivity specifies the percentage of actual faulty tasks (Task Failed) which are correctly classified whereas Specificity is the amount of non-faulty tasks (Task Success) which are correctly identified the relation between these metrics is depicted in confusion matrix.
- **F-measure:** A measure that combines precision and recall.

F-measure = 2*(Precision * recall) / Precision + Recall

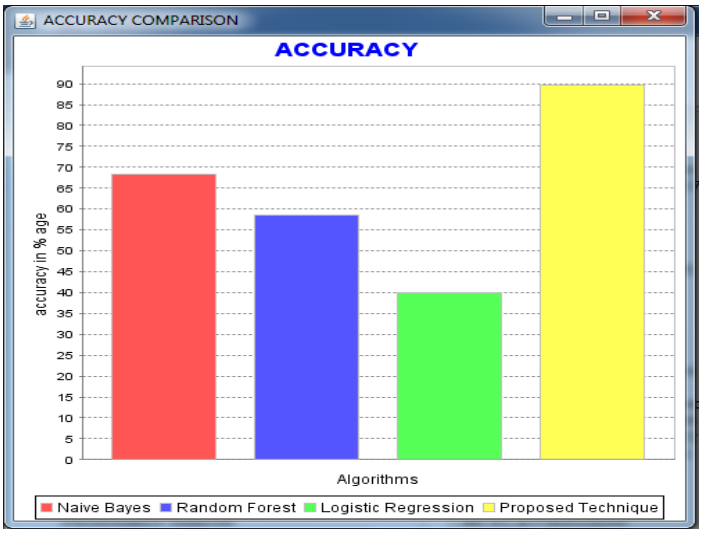

Figure 1: Showing the accuracy comparison of the previous techniques with the proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset using 10 folds Cross Validation Evaluation Model

The comparison graph above shows that the proposed technique performs better for predicting the faults than the previous techniques in terms of accuracy. The accuracy of proposed technique is 89.65 which is greater than all the previous prediction techniques.

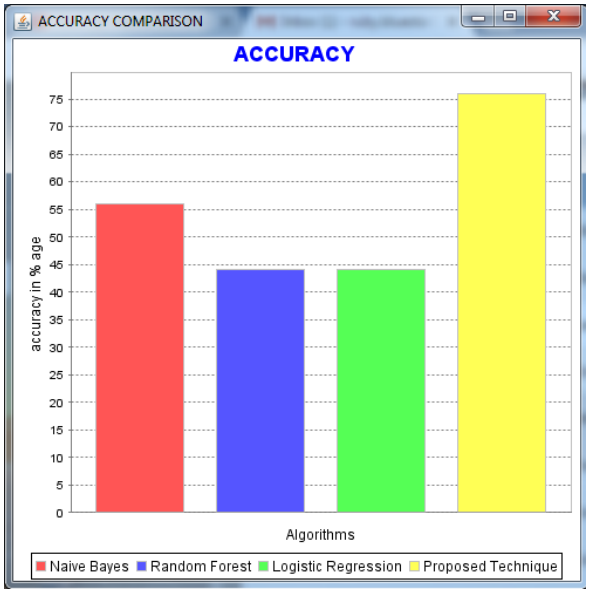

Figure 2: Showing the accuracy comparison of previous techniques with the Proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset using 66-33 Training Testing Percentage split of the data

The comparison graph above shows that the proposed technique performs better for predicting the faults than the previous techniques in terms of accuracy. The accuracy of proposed technique is 76% which is greater than all the previous prediction techniques.

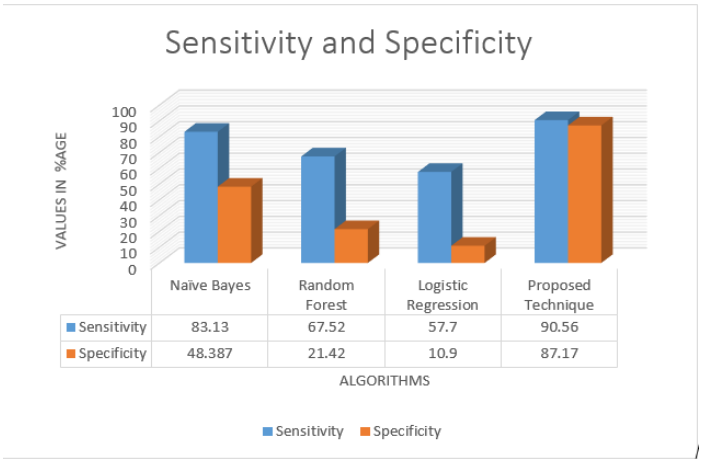

Figure 3: Showing the sensitivity and specificity comparison of previous techniques with the Proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset

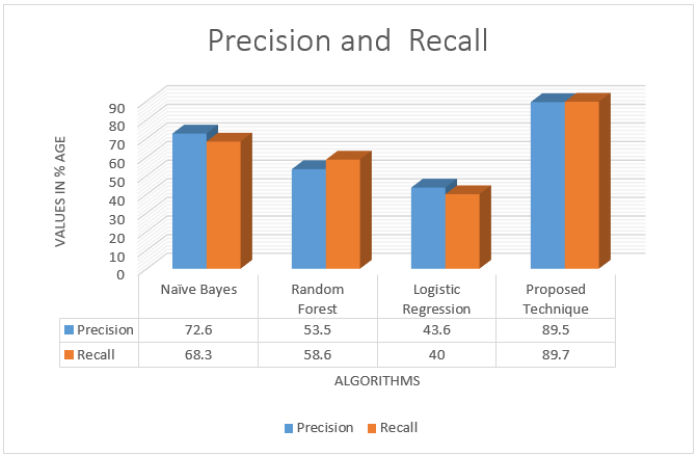

Figure 4: Showing the Precision and recall comparison of previous techniques with the Proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset

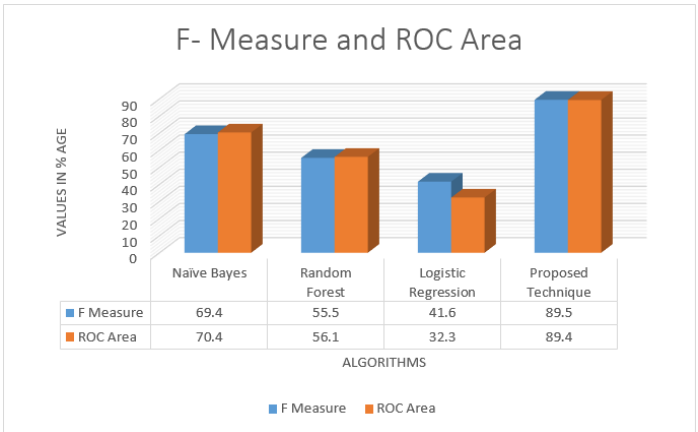

Figure 5: Showing the F Measure and ROC Area comparison of previous techniques with the Proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset

The graphs above show that the proposed technique performs better in all the other parameters also like precision and recall; specificity and sensitivity; F Measure and ROC Area.

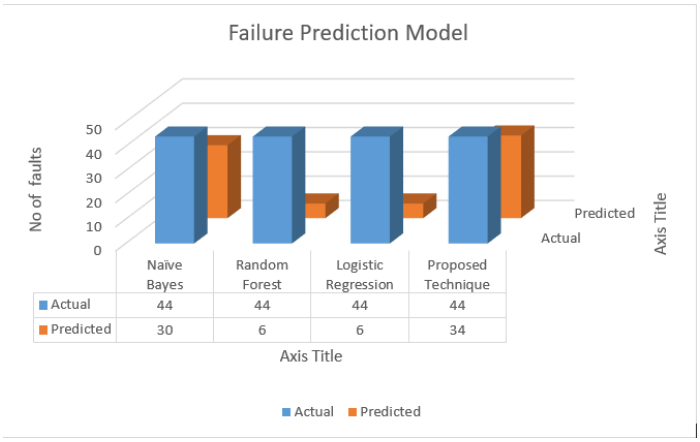

Figure 6: Showing the Actual and predicted faults comparison of previous techniques with the Proposed Ensemble Learning Technique on the generated Montage 100 jobs Dataset

The graphs above show that the proposed technique predicted more no of faults then the previous techniques thus performs better in prediction of fault tolerance model thereby improving the performance accuracy of the proposed fault tolerance prediction model.

CONCLUSION

Cloud environment is dynamic which leads to unexpected system behavior resulting in faults and failures. In order to improve reliability and achieve robustness in cloud computing, failures should be assessed and handled effectively. Fault detection is one of the biggest challenges in making a system fault tolerant this thesis proposes the use of Improved Ensemble Learning technique for predicting the faults in cloud environment using scientific workflows. The faults are first predicted so that suitable fault tolerance technique (pre-emptive migration/ check-pointing) is applied to make the system fault tolerant. The faults will be handled proactively and this will help to resolve the problems associated with fault tolerance techniques.

In future, Ensemble learning can be combined with some more machine learning techniques in order to improve the accuracy of fault prediction model to some more extent. Also, the predicted model can be applied to different scheduling algorithms in workflows in order to enhance the scheduling results. This work may also be extended with the addition of adaptation capabilities to the Pegasus-WMS process, and further analysis to monitor the effectiveness of those diversifications. Additionally, the skills of which workflows are more likely to fail can be utilized to condition dimension granularity.

REFERENCES

[1] Gong, C., Liu, J., Zhang, Q., Chen, H., & Gong, Z. (2010, September). The characteristics of cloud computing. In Parallel Processing Workshops (ICPPW), 2010 39th International Conference on (pp. 275-279). IEEE.

[2] Sun, D. W., Chang, G. R., Gao, S., Jin, L. Z., & Wang, X. W. (2012). Modeling a dynamic data replication strategy to increase system availability in cloud computing environments. Journal of computer science and technology, 27(2), 256-272.

[3] Bala, A., & Chana, I. (2014). Intelligent failure prediction models for scientific workflows. Expert Systems with Applications.

[4] Xiong, N., Vasilakos, A. V., Yang, Y. R., Qiao, C., & Andy, Y. P. (2012). A class of practical self-tuning failure detection schemes for cloud communication networks. IEEE/ACM Transactions on Networking (ToN), submitted.

[5] Meshram, A. D., Sambare, A. S., & Zade, S. D. (2013). Fault Tolerance Model for Reliable Cloud Computing.

[6] Jhawar, R., Piuri, V., & Santambrogio, M. (2013). Fault tolerance management in cloud computing: A system-level perspective. Systems Journal, IEEE, 7(2), 288-297.

[7] Joshi, S. C., & Sivalingam, K. M. (2014). Fault tolerance mechanisms for virtual data center architectures. Photonic Network Communications, 28(2), 154-164.

[8] Deng, J., Huang, S. H., Han, Y. S., & Deng, J. H. (2010, December). Fault-tolerant and reliable computation in cloud computing. In GLOBECOM Workshops (GC Wkshps), 2010 IEEE (pp. 1601- 1605). IEEE.

[9] Nguyen, H., Shen, Z., Tan, Y., & Gu, X. (2013, July). FChain: Toward black-box online fault localization for cloud systems. In Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on (pp. 21-30). IEEE.

[10] Lima, F. R. L. (2004). Improving the quality of service of failure detectors with SNMP and artificial neural networks. In Anais do 22o. Simpósio Brasileiro de Redes de Computadores (pp. 583-586)

